University of Alberta

DETECTION AND DIAGNOSIS OF CONTROL LOOP NONLINEARITIES, VALVE STICTION AND DATA COMPRESSION

by

Md. Ali Ahammad Shoukat Choudhury

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

in

Process Control

Department of Chemical and Materials Engineering

Edmonton, Alberta
Spring 2005

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-08218-6

Your file *Votre référence*
ISBN
Our file *Notre reference*
ISBN

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

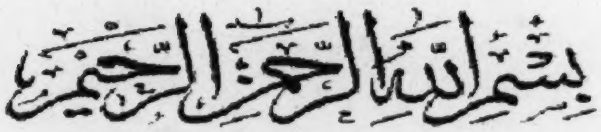
In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada



In the name of Allah, the most beneficent, the most merciful

To
Prophet **Muhammad** (pbuh),
the best person of human history
&
to my parents
for all their kindness to me.

"My choice of Muhammad to lead the list of world's most influential persons may surprise some readers and may be questioned by others, but he was the only man in the history who was supremely successful on both the secular and religious level. It is this unparalleled combination of the secular and religious influence which I feel entitles Muhammad to be considered to be the most influential single figure in human history." - Michael Hart in his book: 'The 100 - A Ranking of the Most Influential Persons in the History'.

"Muhammad was by far the most remarkable man that ever set foot on this earth. He preached a religion, founded a state, built a nation, laid down a moral code, initiated numerous social and political reforms, established a powerful and dynamic society to practice and represent his teachings and completely revolutionized the worlds of human thought and behavior for all times to come." - Sir Bernard Shaw.

"Muhammad is the most successful of all Prophets and religious personalities." - Encyclopedia Britannica.

"Your Lord has decreed that you worship none but Him, and that you be kind to parents. Whether one or both of them attain old age in your life period, say not to them a word of contempt, nor repel them, but address them in terms of honor." - Al-Quran (17:23)

Abstract

The field of controller performance monitoring has received much attention in the engineering research literature. However, the diagnosis of poor control performance remains an open area. Performance diagnosis requires identification of the cause(s) of poor control performance. Poor controller tuning, oscillatory external disturbances, process nonlinearities and valve nonlinearities are the primary causes of poor control performance.

Based on higher order statistical (HOS) theory, two new indices – the Non-Gaussianity Index (*NGI*) and the Non-Linearity Index (*NLI*) – have been developed to detect and quantify signal non-Gaussianity and nonlinearity. These indices, together with specific patterns in the mapping of process output (*pv*) and controller output (*op*), can be used to diagnose the causes of poor control loop performance.

Stiction is the most common problem in spring-diaphragm type valves. A generalized definition of valve stiction based on the investigation of real plant data is proposed in this thesis. A simple two parameter data-driven model of valve stiction is developed. The model is simple, yet powerful enough to properly simulate the complex valve stiction phenomena. Both open and closed loop results have been presented and validated to show the capability of the model.

Conventional invasive methods such as the valve travel test can detect stiction easily. However, they are expensive, time consuming, and tedious to use for examining thousands of valves in a typical process industry. A noninvasive method that can simultaneously detect and quantify control valve stiction is presented. The method requires only routine

operating process data. Over a dozen industrial case studies have demonstrated the wide applicability and practicality of this method as a useful diagnostic aid in troubleshooting poor control performance.

In chemical industrial practice, data are often compressed, for archival purposes, using various techniques. Compression degrades data quality and induces nonlinearity in the data. The issues of data quality degradation and nonlinearity induction due to compression are investigated in this thesis. An automatic method for detection and quantification of the compression present in the archived data has been presented. Compelling and quantitative analyses have been presented to end the practice of process data compression.

Acknowledgements

Surely your Lord is Allah, Who created the heavens and the earth in six periods of time, and He is firm in power; He throws the veil of night over the day, which it pursues incessantly; and He created the sun and the moon and the stars, made subservient by His command; surely His is the creation and the command; blessed is Allah, the Lord of the worlds. (7:54 - Al-Quran).

I express my gratitude to my Lord described above for blessing my life with His countless bounties.

I am indebted to my parents for the support and kindness they provided me throughout my life. I am unable to find a word adequate to express my gratitude to them.

I gratefully acknowledge the outstanding supervision provided to me by Prof. Sirish L. Shah, who was unique in encouraging and inspiring me in various stages of my thesis preparation through his ingenious research directions, friendly discussions, and fatherly affections. This thesis would not have proceeded to completion without his continued stimulating motivation and encouragement.

I am thankful to Prof. Nina F. Thornhill of the University College London (UCL), London, UK, for her excellent guidance, straightforward directions and critical comments. Her suggestions were instrumental in shaping my research work into publishable forms. Many thanks to her for providing me the opportunity to visit UCL and the great city of London.

Financial support in the form of a Canadian International Development Agency (CIDA) scholarship under the auspices of Bangladesh University of Engineering & Technology (BUET)-University of Alberta (UofA) Institutional Linkage Project is gratefully acknowledged. This research project has also been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Matrikon Consulting Inc., and the Alberta Science and Research Authority (ASRA) in the form of an NSERC-Matrikon-ASRA Industrial Research Chair Program at the University of Alberta. Financial support in the form of Andrew Stewart Memorial Graduate Prize, U of A PhD Dissertation

Fellowship, and Captain Thomas Farrel Greenhalgh Memorial Scholarship are also gratefully acknowledged.

I am grateful to Matrikon Inc. for providing me the opportunity to validate my academic research results in an industrial environment. I acknowledge the comments and feedback I received from the people at Matrikon during the various stages of my research project. I would like to especially thank Dave Shook, Warren Mitchell, Murray Faris, Rohit Patwardhan, Jianping Gao, and Fardin Akbaryan.

Special thanks are due to colleagues in the Computer Process Control group for many highly-stimulating discussions in an innovative environment. I am thankful to Dr. Biao Huang, Dr. Frase Forbes, Dr. Scott Meadows, Dr. Arun K. Tangirala, Dr. Shankar Narasimhan, Dr. Sachin Patwardhan, Dr. Weihua Li and Dr. Kamrunnahar for their help. I acknowledge wonderful discussions and friendly time I had with Harigopal Raghavan, Zhengang Han, Vinay Kariwala, Tariq Mannan, Amar bin Halim, Salim Ahmed, Syed Imtiaz, Monjur Murshed, Rumana Sharmin, Nazmul Rahmani, Samina Rahmani, Syeda Razia, Mahbubur Razzaque, Ruouy cheng, and many others.

I would also like to thank the secretaries in the department – Annmarie, Theresa, Leanne, Marilee, and Shona – and the system administrators Bob Barton and Jack Gibeau. I also acknowledge the help of Walter and Richard in the operation of Process Control Labs.

I am thankful to the members of Bangladeshi Students' Association and Muslim Students' Association of U of A for transforming the cold life of Edmonton into a 'cool' place with so many joyous, enjoyable and cheering moments and events.

Last but not least, many thanks are due to my beloved wife, Mahbuba Begum, for her help and moral support, and for the loving and delicious lunches she provided me during my graduate studies at the U of A. The happiness and joy, that our lovely son, Mahdi, brought to our lives, are also remembered.

Table of Contents

1 Introduction 1

1.1 Monitoring Control Loops 1

1.2 Impact of a Faulty Control Valve 3

1.3 Stiction – The Hidden Culprit 3

1.4 Data Quality – Impact of Compression and Quantization 4

1.5 Overview of the Thesis 5

2 Higher Order Statistics: Preliminaries 7

2.1 Introduction 7

2.2 Time Domain Analysis 8

2.2.1 Moments 8

2.2.2 Cumulants 10

2.2.3 The relationship between moments and cumulants 12

2.2.4 Properties of moments and cumulants 13

2.2.5 Moments and cumulants of stationary signals 16

2.3 Spectral Analysis 16

2.3.1 Power spectrum, $n = 2$ 17

2.3.2 Bispectrum, $n = 3$ 18

2.3.3 Estimation of the bispectrum 19

2.3.4 Properties of estimators and asymptotic behaviour 22

2.3.5 Bicoherence or normalized bispectrum 23

2.3.6 Properties of bispectrum and bicoherence 25

2.4 Bispectrum or Bicoherence Estimation Issues 27

2.4.1 Choice of window function 29

2.4.2 Choice of data length, segment length, and fourier transform length 29

2.5 Conclusions 31

3 Diagnosis of Poor Control Performance 32
3.1 Introduction . . . 33
3.2 Problem Formulation . . . 34
3.3 Test of Gaussianity and Linearity of a Signal . . . 35
3.4 Illustrative Examples . . . 40
3.4.1 Bicoherence of a linear and nonlinear signal . . . 40
3.4.2 Bicoherence of a nonlinear sinusoid signal with noise . . . 41
3.5 Investigation of Control Loop Nonlinearities . . . 45
3.5.1 Process nonlinearity . . . 45
3.5.2 Nonlinear valve characteristic . . . 48
3.5.3 Nonlinear faults in valve . . . 52
3.5.4 Nonlinear disturbances . . . 53
3.6 Simulation Example to Diagnose the Causes of Poor Performance . . . 53
3.6.1 Well tuned controller . . . 54
3.6.2 Tightly tuned controller . . . 54
3.6.3 Presence of an external oscillatory disturbance . . . 55
3.6.4 Presence of stiction . . . 56
3.7 Industrial Case Studies . . . 56
3.7.1 Stiction in a furnace dryer temperature control valve . . . 57
3.7.2 Valve Saturation . . . 58
3.7.3 Valve problems in some flow control loops . . . 59
3.8 Conclusions . . . 61

4 Modelling Control Valve Stiction 62
4.1 Introduction . . . 63
4.2 What is Stiction? . . . 65
4.2.1 Definition of terms relating to valve nonlinearity . . . 65
4.2.2 Discussion of the term "Stiction" . . . 67
4.2.3 A proposal for a definition of stiction . . . 68
4.3 Practical Examples of Valve Stiction . . . 70
4.4 A Physical Model of Valve Friction . . . 74
4.4.1 Model formulation . . . 74
4.4.2 Valve simulation . . . 77
4.5 Data Driven Model of Valve Stiction . . . 80
4.5.1 Model formulation . . . 81

4.5.2 Open loop response of the model under a sinusoidal input 84
4.5.3 Closed loop behaviour of the model 86
4.6 Describing Function Analysis 90
4.6.1 Introduction 90
4.6.2 An expression for the describing function 90
4.6.3 Asymptotes of the describing function 91
4.6.4 Insights gained from the describing function 91
4.7 Conclusion 93
4.A Appendix – Derivation of Describing Function 94
4.A.1 Expression for the output of the non-linearity 94
4.A.2 Evaluation of the fundamental Fourier component 95
4.A.3 Evaluation of limiting cases 96

5 Automatic Detection and Quantification of Control Valve Stiction **97**
5.1 Introduction 98
5.2 Detection of Stiction in Control Valves 99
5.2.1 Detection of loop nonlinearity 100
5.2.2 Use of *pv-op* plot 103
5.3 Quantifying Stiction 105
5.3.1 Clustering technique 105
5.3.2 Using a fitted ellipse 106
5.4 An Illustrative Example 106
5.5 Automatic Detection and Quantification of Stiction 109
5.6 Simulation Results 111
5.6.1 Diagnosis of stiction 111
5.6.2 Diagnosis of a sinusoidal disturbance 114
5.6.3 Diagnosis of root cause of a propagated disturbance 115
5.7 Industrial Case Studies 118
5.7.1 Loop 1 : A level loop 118
5.7.2 Loop 2: A linear level control loop 120
5.7.3 Loop 3: A flow control loop 121
5.7.4 Loop 4: A temperature control loop 122
5.7.5 Loop 5: A pressure control loop 123
5.7.6 Loop 6: A composition control loop 124
5.7.7 Loop 7: A cascaded flow control loop 125

5.8 Online Compensation for Stiction 126
5.9 Conclusions 126
5.A Appendix Clustering Techniques 127
5.A.1 C-means clustering 127
5.A.2 Fuzzy C-means clustering 127
5.B Appendix Fitting an ellipse 128

6 Impact of Data Compression and Quantization on Data-Driven Process Analyses 130
6.1 Introduction 131
6.2 Motivating example 132
6.3 Methods 133
6.3.1 Overview of data compression 133
6.3.2 Direct methods 133
6.3.3 Transform methods 137
6.3.4 Selected compression method – swinging door compression 137
6.4 Measures of Data Quality 138
6.4.1 Statistical properties 138
6.4.2 Non-linearity measure 138
6.4.3 Performance index (Harris) measures 139
6.5 Process Data for Compression Comparison 140
6.5.1 Industrial example 1 140
6.5.2 Industrial example 2 142
6.6 Results and Discussions for Industrial Example 1 144
6.6.1 Visual observations 144
6.6.2 Statistical properties 144
6.6.3 Nonlinearity assessment 146
6.6.4 Performance (Harris) index 146
6.7 Results and Discussions for Industrial Example 2 147
6.7.1 Visual observations 147
6.7.2 Statistical properties 148
6.7.3 Nonlinearity assessment 149
6.7.4 Performance (Harris) index 149
6.8 Summary of Data Quality Measures 150
6.9 Automated Detection of Compression 151

6.9.1 Motivation 151
6.9.2 Compression detection procedure 151
6.9.3 Implementation considerations 153
6.10 Final Recommendations Concerning Compression 157
6.10.1 Discussion on the motivating example 157
6.10.2 A proposal for harmless storing of data 157
6.11 Quantization 158
6.12 Conclusions 160

7 Conclusions **161**
7.1 Contributions of This Thesis 161
7.2 Recommendations for Further Work 162

Bibliography **164**

List of Tables

3.1 QPC relations for the output signal, y 42

4.1 Nominal values used for physical valve simulation 76
4.2 Friction values used in simulation of physical valve model 77
4.3 Process models, controllers, and data-driven stiction model parameters. . . 86

5.1 Numerical Results for the Industrial Loops Analyses 118

List of Figures

2.1 *Principal domain of the bispectrum.* 19
2.2 *Information flow diagram for the direct method of bispectrum estimation.* . 21
2.3 *Bicoherence plot for the QPC example.* 28
2.4 *The effect of signal to noise ratio and DFT length on bispectral estimation for a quadratically coupled harmonic signal. Larger DFT length and higher SNR provide better estimates.* 30

3.1 *Global multi-industry performance demographics.* 33
3.2 *A typical control loop under feedback control.* 34
3.3 *Rule based decision flow diagram.* 39
3.4 *Results for y_{linear} (top) and $y_{nonlinear}$ (bottom).* 41
3.5 *HOS analysis results for the linear and nonlinear sinusoid signals, case 1 - mild nonlinearity.* 43
3.6 *HOS analysis results for case 2 - strong nonlinearity.* 44
3.7 *Level control of a spherical tank.* 46
3.8 *Results of the analysis of open loop data for the spherical tank system.* . . . 46
3.9 *Squared bicoherence for small excitation of spherical tank level.* 47
3.10 *Results of the analysis of closed loop data for the spherical tank system.* . . 48
3.11 *Squared bicoherence for small excitation of the spherical tank level under feedback control. The flatness of the plot shows the linearity of the process.* 48
3.12 *Inherent characteristics for linear, equal percentage and square-root valve.* 49
3.13 *Block diagram of a simple SISO process with stiction nonlinearity in the valve.* 50
3.14 *Bicoherence plot for an equal percentage valve. Left - Valve travels less than 25% of its full span. Right- Valve travels more than 25% of its full span.* 51
3.15 *Input-output characteristics of valve saturation.* 53

3.16 *Results of the simulation example. Bicoherence correctly detects the first three cases as linear and the last case as nonlinear. The pv-op plot for the stiction case shows elliptic patterns.* 55
3.17 *Analysis of time series data from an industrial temperature control loop.* . . 57
3.18 *Results of the analysis of an industrial loop with valve saturation.* 58
3.19 *Analysis of flow loop 1 data before (April - left) and after the (July - right) plant maintenance shutdown period.* 59
3.20 *Analysis of flow loop 2 data before (April - top) and after (July - bottom) the plant maintenance shutdown period.* 60

4.1 *Input-Output behaviour of hysteresis, deadband, and deadzone (redrawn from ANSI/ISA-S51.1-1979).* 66
4.2 *Typical input-output behaviour of a sticky valve.* 69
4.3 *Flow control cascaded to level control in an industrial setting, the line with circles is pv and mv, the thin line is op.* 71
4.4 *Data from a flow loop in a refinery, time trend of pv and op (left) - the line with circles is pv and the thin line is op, and the pv-op plot (right).* 72
4.5 *Data from a flow loop in a refinery, time trend of pv and op (top left) - the line with circles is pv and the thin line is op, the pv-op plot(top right), time trend of pv and sp (bottom left), line with circles is pv and thin line is sp, and the pv-sp plot(bottom right).* 72
4.6 *Industrial dryer temperature control loop data, lines with circles are pv and flowrate, the thin line is op (the bottom left panel).* 73
4.7 *A cross-sectional diagram of a pneumatic control valve.* 75
4.8 *Friction characteristic plot.* 76
4.9 *Open loop response of mechanistic model. The amplitude of the sinusoidal input is 10 cm in each case.* 78
4.10 *Closed loop response of mechanistic model.* 79
4.11 *Signal and logic flow chart of the data-driven stiction model.* 82
4.12 *Open loop simulation results of the data-driven stiction model.* 85
4.13 *Closed loop simulation results of a concentration loop in presence of the data-driven stiction model.* 88
4.14 *Closed loop simulation results of a level loop in presence of the data-driven stiction model.* 89

4.15 *Input (thin line) and output (heavy line) time trends for the limiting case as $X_m = S/2$. Left panel: Slip-jump only with $S = J$. Right panel: Deadband only with $J = 0$. The output in the right plot has been magnified for visualization; its amplitude becomes zero as X_m approaches $S/2$.* 92
4.16 *Graphical solutions for limit cycle oscillations. Left panel: Composition control loop. Right panel: level control loop. Dotted lines are the $-1/N$ curves and the solid line is the frequency response function.* 93
4A.1 *Describing function analysis for stiction nonlinearity.* 94

5.1 *Decision flow diagram of the methodology for the detection and diagnosis of loop nonlinearity.* 102
5.2 *Data analysis results for an industrial level control loop. The biocherence plot shows large peaks indicating nonlinearities in the loop. The subplots (c), (d) and (e) demonstrate c-means and fuzzy c-means method of quantifying stiction. The ellipse fitting technique of quantifying stiction is shown in the subplot (f). Approximately, 11% stiction was estimated in this control valve.* 107
5.3 *Valve Position (mv) versus Controller Output (op) plot. This plot confirms that the amount of stiction was correctly estimated in Figure 5.2.* 108
5.4 *Decision flow diagram of the methodology for the detection and quantification of valve stiction.* 110
5.5 *Block diagram of a simple SISO process with stiction nonlinearity in the valve.* 112
5.6 *Results of detection and quantification of stiction in a simulated data set.* . . . 113
5.7 *Stiction detection results for a simulated data set when an external sinusoidal oscillatory disturbance plus noise are entering the loop. The cross-correlation method detects stiction, even though there was no stiction (plot (b)). The bicoherence based method correctly shows that there is no nonlinearity this loop (subplot(c)).* 114
5.8 *Simulation Block diagram for stiction induced oscillation propagation study.* 116
5.9 *Analysis results for root cause diagnosis of a propagated disturbance. Result for the composition loops with a sticky valve (left). Results for the level loop affected by the disturbance from the composition loop (right).* . . . 117

5.10 *Stiction detection and quantification results for an industrial level control loop data. Significantly large peaks in the bicoherence plot detects nonlinearity in the loop. The presence of ellipse confirms stiction. The width of the ellipse or c-means clustering method quantifies stiction.* 119
5.11 *Results of the analysis of another industrial level control loop data. The flatness of the bicoherence plot shows the linearity of this level control valve, which was confirmed in the valve position versus valve input plot (subplot (c)).* 120
5.12 *Results of the analysis of a refinery flow control loop data. Significant peaks in the bicoherence plot show the nonlinearities in the loop. The presence of ellipse in pv-op plot confirms stiction.* 121
5.13 *Results of stiction detection and quantification in an industrial furnace dryer temperature control loop. Approximately 1.1% stiction was detected in the natural gas flow control valve.* 122
5.14 *Results of the analysis of a refinery pressure control loop data. Approximately 11% stiction was present in the control valve.* 123
5.15 *Results of the analysis of an industrial composition control loop data. Approximately 1% stiction was present in the control valve.* 124
5.16 *Results of the analysis of an industrial cascaded flow loop data. The bicoherence plot shows the nonlinearity present in this loop. The absence of an ellipse in the pv-op plot (subplot (c)) indicates that this nonlinearity is not due to valve stiction.* 125

6.1 *An industrial data set with compression in some tags. Time trends are mean centered and normalized. CF as shown on the right column is an estimate of the compression factor (as defined in Section 6.9.2).* 132
6.2 *Box car algorithm for data compression (from Hale and Sellar (1981)).* . . 134
6.3 *Backward slope compression algorithm (from Hale and Sellar (1981)).* . . . 135
6.4 *Combined Box car and backward slope data compression algorithm (from Hale and Sellar (1981)).* 135
6.5 *Swinging door algorithm for data compression.* 136
6.6 *Time trends of original data (open circles for data set 1, squares for data set 2, and diamonds for data set 3) and reconstructed data with compression factor 10 (lines).* 141

6.7 *Power spectrum of the original data and reconstructed data with compression factor 10, original spectra (dotted line) and spectra of reconstructed signals (solid line).* 142
6.8 *Time trends of original data (open circles for flow loop data, squares for temperature loop data, and diamonds for pressure loop data) and reconstructed data with compression factor 10 (lines).* 143
6.9 *Power spectrum of the original data and reconstructed data with compression factor 10, original spectra (dotted line) and spectra of reconstructed signals (solid line).* 143
6.10 *Statistical measures as a function of CF for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).* 145
6.11 *Non-Gaussianity Index, Nonlinearity index, and Harris index as a function of compression factor for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).* 147
6.12 *Statistical measures as a function of CF for Flow loop (circles), temperature loop (squares) and pressure loop data (diamonds).* 148
6.13 *Non-Gaussianity Index, Nonlinearity Index, and Harris index as a function of compression factor for flow loop data (circles), temperature loop data (squares) and pressure loop data (diamonds).* 149
6.14 *Results from the compression estimation algorithm for data sets 1 to 3, data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).* 152
6.15 *Results from the compression estimation algorithm, flow loop (circles), temperature loop (squares) and pressure loop (diamonds).* 153
6.16 *Left panel shows the second derivative of the flow trend, Right panel shows the distribution of the second derivative.* 156
6.17 *The variation of central bin size with number of bins.* 156
6.18 *An algorithm for storing of data with little loss of information.* 157
6.19 *An example of heavily quantized data representing a process variable.* . . . 158
6.20 *An industrial data set where some of the variables were heavily quantized.* . 159

1 Introduction

Modern process industries are increasingly automated to achieve objectives such as maintaining world class product quality, reducing operating and maintenance costs, enhancing operator safety, meeting environmental and occupational health regulations, optimizing resource management, and increasing profitability from resources. One of the manifestations of this increased automation is the increasing number of controllers and control loops in process industries. A process plant may have from a few hundred control loops to several thousands depending on the complexity of the plant. Clearly, monitoring and assessing the performance of these control loops are crucial to process industries.

1.1 Monitoring Control Loops

Satisfactory control performance is important to ensure high product quality and low product cost in chemical plants. The controllers may initially be tuned properly for high performance, but normal "wear and tear" and changing operating conditions may contribute to the deterioration of the performance of such controllers. Common causes of sub-optimal process operation include (1) changes in operating regimes of the plant, (2) changes in operating conditions, (3) changes in plant production rates due to market conditions, (4) partial modification of the plant, (5) wear and tear of mechanical equipment such as control valves and sensors, (6) inadequate instrument maintenance and (7) lack of periodic tunings

of the controllers.

Poor loop performance can sometimes be easily detected by the operator, but often the problem may propagate to other loops, thereby making it difficult to detect the root cause. Regular performance monitoring of control loops can enable one to detect poorly-performing loops and provide diagnostic aids to repair the problematic loops. Desborough and Harris (1993), Harris (1989) and Stanfelj *et al.* (1993) discussed several methods for controller performance assessment with minimum variance and settling time benchmarks from routine operating data. Modern process industries are increasingly interested in loop performance monitoring for the following reasons:

- Poor performance of controllers adversely affects key performance indicators (KPI) of the plants. The economic benefits resulting from performance assessment are difficult to quantify on a loop-by-loop basis because each problem loop contributes in a complicated way to the overall process performance. Finding and fixing problem loops throughout a plant can improve product quality, reduce product property variability, lower operating costs and increase production rate (Paulonis and Cox, 2003). Even a 1% improvement either in energy efficiency or reduced product variability saves hundreds of millions of dollars for process industries (Desborough and Miller, 2002).

- Improvement in the performance of the controllers allows engineers to easily cope with the increasingly stringent environmental regulations imposed on chemical plants.

- Advanced Process Control (APC) strategies such as Model Predictive Control (MPC) are increasingly used to meet rapidly-changing market conditions. In many cases, the performance of these MPCs depends in turn on satisfactory performance of lower level regulatory Proportional-Integral-Derivative (PID) controllers.

- Through regular performance monitoring and preventative maintenance of the controllers and actuators (valves), in many cases it is possible to avoid unscheduled shutdown of the plant. Every unscheduled shut-off or downtime of a plant takes a costly toll on the maintenance budget of the plant.

1.2 Impact of a Faulty Control Valve

The control valve is typically the actuator for most of process control loops. It is the only moving part in the loop. It serves as the 'work-horse' that implements the control decision. If the control valve malfunctions, the performance of the loop is likely to deteriorate, no matter how good the controller is. Commonly encountered control valve problems include stiction, hysteresis, backlash, deadband and saturation. Because of these problems, the valve output may be oscillatory which in turn may cause oscillations in many process variables. Other reasons for an oscillatory control loop include poor controller tuning, improper process design, poor control system configuration, and oscillatory disturbances (Bialkowski, 1992; Ender, 1993; Miao and Seborg, 1999). Bialkowski (1992) reported that about 30% of the loops are oscillatory due to control valve problems. Oscillatory variables are one of the main causes of poor performance of control loops. A key challenge is to find the root cause of distributed oscillations in chemical plants (Qin, 1998; Thornhill *et al.*, 2003*a*; Thornhill *et al.*, 2003*b*). The presence of oscillations in a control loop increases the variability of the process variables, thus causing inferior quality products, larger rejection rates, increased energy consumption, decreased average throughput and reduced profitability. Oscillations can cause a valve to wear out much earlier than the life period for which it was originally designed. Oscillations increase operating costs roughly in proportion to the deviation (Shinskey, 1990). Detection and diagnosis of the causes of oscillations in process operation are important because a plant operating close to product quality limit is more profitable than a plant that has to back away because of variations in the product (Martin *et al.*, 1991). The presence of a faulty valve in a control loop hinders the achievement of good performance of the control loops. The economic impact of poor performance of control loops has been discussed in the previous section.

1.3 Stiction – The Hidden Culprit

Stiction in control valves is one of the long-standing and common problems in the process industry. Stiction hinders the proper movement of the valve, which in turn degrades the performance of the control loop. Stiction is one of the main root causes for initiating oscillation(s) in a control loop. The oscillation may propagate throughout the plant. The scenario becomes even more complex if there is a recycle stream in the plant. Indeed recycles are commonplace in most chemical industries. The detection and quantification of stiction is the main focus of this study. There are two categories

of methods to detect stiction: invasive and non-invasive. The invasive method requires stroking/travelling the valve over its full travel span when in-service or out of service. Because it is neither feasible nor cost-effective to invasively test hundreds of valves at a plant site, the non-invasive method is preferred to the invasive one. Horch's cross-correlation method (Horch, 1999; Horch *et al.*, 2000; Horch, 2000) is one of the popular non-invasive methods for detecting stiction reported to date. Horch's method detects stiction with the use of the cross-correlation function between *pv* and *op*. His method is not applicable to processes containing an integrator (e.g., a level control loop), or for loops carrying compressible media (e.g., steam or air). The Horch method is mainly useful for flow control loops. However, even for flow control loops it sometime produces inconclusive results (Desborough and Miller, 2002). Also, if there is a sinusoidal disturbance entering the control loop, the method falsely detects stiction in the control valve (Choudhury *et al.*, 2002). More details will be provided in Chapter 5. Moreover, none of the existing methods can quantify stiction. In this thesis a new method to detect and quantify stiction has been developed. The method is based on higher order statistics and applicable to all types of control loops.

1.4 Data Quality – Impact of Compression and Quantization

Data quality lies at the heart of all data driven process analyses. It plays a central role in determining the credibility of all data analyses. Data quality issues include determination or quantification of randomness, measures of information content, stationarity, compression factors, and quantization levels. In the analysis of chemical process data, compression and quantization are frequently encountered problems.

In chemical industrial practice, data are often compressed using various data compression techniques such as box car, backward slope, swinging door and wavelet compression algorithms before storing them in the historian (Kennedy, 1993; Aspentech, 2001; Watson, 1993; Hale and Sellars, 1981; Bristol, 1990). Historical data are an invaluable source of information. However, compression degrades data quality and induces nonlinearity in the data (Watson *et al.*, 1998). The issues of data quality degradation and nonlinearity induction due to compression should be properly dealt with prior to processing the data for performance monitoring. Until recently, there were serious concerns regarding compressed data analysis, but there has been no systematic study of the impact

of compression on process data. An automatic method for detection and quantification of compression present in archived data is an invaluable tool during the preprocessing stage of any data analysis. A method of quantifying the degree of compression in archived or historical data is developed in this thesis (Thornhill *et al.*, 2004). Higher order statistical techniques were also found to be useful in the analysis of compression induced nonlinearity in data.

Quantization is a problem that arises during the analog to digital conversion of data (Smith, 1998; Ifeachor and Jervis, 1993). Modern control systems involve computers and digital equipments. Because computers cannot read an analog signal, all analog signals must be converted into digital signals. For digitization, two steps are required: sampling and quantization. Sampling is only the first phase of acquiring data into a computer. Computational processing quantizes data, i.e., analog values are converted into digital form. The number of bits determines the precision of the digitized data. Old A/D converters have lower numbers of bits, i.e., lower resolution. These may introduce a significant amount of quantization errors. A Significant amount of quantization errors may produce oscillations (Horch, 2000) in process variables. Sometime the quantization errors are too large to use the data for any practical purposes.

1.5 Overview of the Thesis

This thesis is divided into seven chapters including this introductory chapter. Most of the chapters are independent of each other and can be read in any order according to the reader's preference. A summary of each chapter follows:

- Chapter 1 discusses the importance of the monitoring of control loop performance, the impact of a faulty valve (especially a sticky valve) on the performance of a control loop, and the issues of data quality degradation due to compression and quantization.

- Higher order statistics plays a central role in this thesis in detecting the presence of nonlinearities in control loops. A tutorial introduction to this subject matter is presented in chapter 2.

- Chapter 3 describes the development and application of higher order statistics based tools used for diagnosing poor loop performance. It also discusses several simulation and industrial case studies.

- Stiction, the most commonly encountered control valve problem, is defined, discussed and modelled in chapter 4. A proper definition of stiction is proposed in this chapter. New insights are gained from the describing function analysis of the newly defined stiction model.

- When there are thousands of control loops in a process industry, it is almost impossible to check each valve manually. The challenge is to develop an automated method that requires minimal human input to monitor the health of each control valve. Chapter 5 discusses the methodology to automatically detect and quantify stiction in control valves.

- Historical data are an invaluable source of information. In chemical industrial practice, data are often compressed using various techniques before being stored or archived in historians. Compression degrades data quality and induces nonlinearity into the data. Chapter 6 focuses on the problems of data quality degradation and nonlinearity induction due to compression. It describes an automatic method to detect and quantify compression present in the archived data. A method for detection and quantification of quantization in process data is also presented.

- Chapter 7 concludes the thesis with a summary of its main contributions. It also provides directions for future work in this area.

2

Higher Order Statistics: Preliminaries

A tutorial introduction to Higher Order Statistics (HOS) and its relation to conventional Second Order Statistics (SOS) are presented in this chapter. The first and second order statistics, for example, mean, variance, autocorrelation, and power spectrum, are popular signal processing tools and are used extensively in data analysis. Such second order statistics are sufficient for describing linear and Gaussian processes. In practice, there are many situations where the process deviates from Gaussianity and linearity — for example, when it exhibits nonlinear behavior. These type of processes can conveniently be studied using Higher Order Statistics (HOS).

2.1 Introduction

A signal is a variable that carries some kind of information that can be conveyed, displayed or manipulated. Signals can be obtained from numerous sources - for example, speech, sound, music, images, chemical processes, radar, seismic surveys, computer

simulations, robots, etc. Useful information is often hidden in a signal. The objective in a signal processing problem is to process a signal or a finite number of data samples to extract important information that may be "hidden" therein (Smith, 1998; Ifeachor and Jervis, 1993). In most cases, data are available as variables sampled at regular intervals of time. Digital signal processing continues to play an important role in the industrial revolution of high technology. Digital signal processing — combined with the ideas and methodologies from various branches of engineering (computer engineering, chemical engineering, electrical and electronics engineering), statistics, numerical analysis and computer science — has become a very useful tool for data analysis and is being used in many industrial processes. The data are usually analyzed either in the time domain or in the frequency domain.

2.2 Time Domain Analysis

Time domain or time series data are a good source of information. Many statistical measures (e.g., moments, cumulants, auto-correlation, cross-correlation) have been developed to measure the temporal signal characteristics of such data. Almost all types of data are usually collected as samples at regular intervals of time. In statistical analysis, it is often assumed that the time series or the signal is stationary. This assumption holds for subsequent definitions and analyses.

2.2.1 Moments

A generating function of a random variable is an expected value of a certain transformation of the variable. All generating functions have three important properties:

1. The generating function completely determines the distribution.
2. The generating function of a sum of two independent variables is the product of the independent generating functions.
3. The moments of the random variable distribution can be obtained from the derivatives of the generating function.

As the term implies the moment generating function should be able to generate all moments. For any random variable x, the moment generating function can be defined

as the expectation of the transformation, e^{tx}, where $t \in R$, i.e.,

$$M_x(t) = E[e^{tx}] \tag{2.1}$$

Moments can be obtained from the coefficients of the Taylor's series expansion of the moment generating function about the origin,

$$\begin{aligned} M_x(t) &= M_x(t)\,|_{t=0} + \left.\frac{\partial M_x(t)}{\partial t}\right|_{t=0}(t-0) + \frac{1}{2!}\left.\frac{\partial^2 M_x(t)}{\partial t^2}\right|_{t=0}(t-0)^2 + \\ &\quad \frac{1}{3!}\left.\frac{\partial^3 M_x(t)}{\partial t^3}\right|_{t=0}(t-0)^3 + \cdots \\ &= M_x(t)\,|_{t=0} + \left.\frac{\partial M_x(t)}{\partial t}\right|_{t=0} t + \frac{1}{2!}\left.\frac{\partial^2 M_x(t)}{\partial t^2}\right|_{t=0} t^2 + \\ &\quad \frac{1}{3!}\left.\frac{\partial^3 M_x(t)}{\partial t^3}\right|_{t=0} t^3 + \cdots \end{aligned} \tag{2.2}$$

From the right side of the Equation 2.2, the first derivative gives the first order moment:

$$\begin{aligned} m_1 &= \left.\frac{\partial M_x(t)}{\partial t}\right|_{t=0} \\ &= \left.\frac{\partial E[e^{tx}]}{\partial t}\right|_{t=0} \\ &= E[xe^{tx}]\,|_{t=0} \\ &= E[x] \end{aligned} \tag{2.3}$$

The second derivative gives the second order moment:

$$m_2 = \left.\frac{\partial^2 M_x(t)}{\partial t^2}\right|_{t=0} = \left.\frac{\partial^2 E[e^{tx}]}{\partial t^2}\right|_{t=0} = E[x^2 e^{tx}]\,|_{t=0} = E[x^2] \tag{2.4}$$

Similarly, the third derivative gives the third order moment:

$$m_3 = E[x^3] \tag{2.5}$$

and so on. Therefore, we can rewrite the moment generating function as

$$\begin{aligned} M_x(t) &= E[e^{tx}] \\ &= E[1 + tx + \frac{t^2}{2!}x^2 + \frac{t^3}{3!}x^3 + \cdots] \\ &= 1 + tE[x] + \frac{t^2}{2!}E[x^2] + \frac{t^3}{3!}E[x^3] + \cdots \\ &= 1 + tm_1 + \frac{t^2}{2!}m_2 + \frac{t^3}{3!}m_3 + \cdots \end{aligned} \tag{2.6}$$

The first order moment is the mean ($\mu = E(x)$) of the data series, $x(k)$, which provides a measure of the location or the center of gravity of the probability density function (pdf) for an ergodic signal. The second order moment is the variance of the data series and gives the spread of the pdf, while the third order moment provides a measure of the skewness of the distribution and the fourth order moment provides a measure of the flatness of the distribution. Alternatively, for any zero-mean data series, $x(k)$, the 2^{nd}, 3^{rd}, ..., n^{th} order moment of the data series can be defined as:

$$\begin{aligned} m_2(\tau_1) &\triangleq E[x(k)x(k+\tau_1)] \\ m_3(\tau_1,\tau_2) &\triangleq E[x(k)x(k+\tau_1)x(k+\tau_2)] \\ &\vdots \quad\quad \vdots \\ m_n(\tau_1,\tau_2,\cdots\tau_{n-1}) &\triangleq E[x(k)x(k+\tau_1)x(k+\tau_2)\cdots x(k+\tau_{n-1})] \end{aligned} \tag{2.7}$$

The variance, σ^2, can be obtained from $m_2(0)$; skewness from $m_3(0,0)$; and so on.

2.2.2 Cumulants

Cumulants are another set of statistical measures that can be used instead of moments because of their excellent noise suppressing properties. The cumulant generating function is defined as the logarithmic of the moment generating function. That is, for a random variable x, the cumulant generating function is

$$C_x(t) \triangleq \ln(M_x(t)) \tag{2.8}$$

Just as moments are derived from the Taylor's series expansion of the moment generating function, cumulants can be derived from the Taylor's series expansion of the cumulant generating function:

$$\begin{aligned} C_x(t) &= C_x(t)\,|_{t=0} + \left.\frac{\partial C_x(t)}{\partial t}\right|_{t=0}(t-0) + \frac{1}{2!}\left.\frac{\partial^2 C_x(t)}{\partial t^2}\right|_{t=0}(t-0)^2 + \\ &\quad \frac{1}{3!}\left.\frac{\partial^3 C_x(t)}{\partial t^3}\right|_{t=0}(t-0)^3 + \cdots \\ &= C_x(t)\,|_{t=0} + \left.\frac{\partial C_x(t)}{\partial t}\right|_{t=0} t + \frac{1}{2!}\left.\frac{\partial^2 C_x(t)}{\partial t^2}\right|_{t=0} t^2 + \\ &\quad \frac{1}{3!}\left.\frac{\partial^3 C_x(t)}{\partial t^3}\right|_{t=0} t^3 + \cdots \end{aligned} \tag{2.9}$$

The first order cumulant is given by:

$$\begin{aligned} c_1 &= \left.\frac{\partial C_x(t)}{\partial t}\right|_{t=0} \\ &= \left.\frac{\partial \ln[E[e^{tx}]]}{\partial t}\right|_{t=0} \\ &= \left.\frac{\partial \ln[1+tm_1+\frac{t^2}{2!}m_2+\frac{t^3}{3!}m_3+\cdots]}{\partial t}\right|_{t=0} \\ &= \left.\frac{1}{[1+tm_1+\frac{t^2}{2!}m_2+\frac{t^3}{3!}m_3+\cdots]}[m_1+\frac{2t}{2!}m_2+\frac{3t^2}{3!}m_3+\cdots]\right|_{t=0} \\ &= m_1 \end{aligned} \tag{2.10}$$

The second order cumulant is:

$$\begin{aligned} c_2 &= \left.\frac{\partial^2 C_x(t)}{\partial t^2}\right|_{t=0} \\ &= \left.\frac{\partial^2 \ln[E[e^{tx}]]}{\partial t^2}\right|_{t=0} \\ &= \left.\frac{\partial^2 \ln[1+tm_1+\frac{t^2}{2!}m_2+\frac{t^3}{3!}m_3+\cdots]}{\partial t^2}\right|_{t=0} \\ &= \left.\frac{\partial}{\partial t}\left[\frac{1}{[1+tm_1+\frac{t^2}{2!}m_2+\frac{t^3}{3!}m_3+\cdots]}[m_1+\frac{2t}{2!}m_2+\frac{3t^2}{3!}m_3+\cdots]\right]\right|_{t=0} \\ &= \left.\left[\frac{-1}{[1+tm_1+\frac{t^2}{2!}m_2+\frac{t^3}{3!}m_3+\cdots]^2}[m_1+\frac{2t}{2!}m_2+\frac{3t^2}{3!}m_3+\cdots]^2\right]\right|_{t=0} \\ &\quad + \left.\left[\frac{1}{[1+tm_1+\frac{t^2}{2!}m_2+\frac{t^3}{3!}m_3+\cdots]}[0+m_2+\frac{6t}{3!}m_3+\cdots]\right]\right|_{t=0} \\ &= -m_1^2+m_2 \\ &= m_2-m_1^2 \end{aligned} \tag{2.11}$$

Similarly, the third order cumulant is:

$$\begin{aligned} c_3 &= \left.\frac{\partial^3 C_x(t)}{\partial t^3}\right|_{t=0} \\ &= m_3-3m_2m_1+2m_1^3 \end{aligned} \tag{2.12}$$

The fourth order cumulant

$$\begin{aligned} c_4 &= \left.\frac{\partial^4 C_x(t)}{\partial t^4}\right|_{t=0} \\ &= m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4 \end{aligned} \tag{2.13}$$

Note that all of the above cumulants are about the origin. For any other point (τ), the 2^{nd}, 3^{rd}, ..., n^{th} order cumulant of any random data series, $x(k)$, can be rewritten as:

$$\begin{aligned} c_2(\tau_1) &\triangleq cum[x(k), x(k+\tau_1)] \\ c_3(\tau_1, \tau_2) &\triangleq cum[x(k), x(k+\tau_1), x(k+\tau_2)] \\ &\vdots \quad\quad \vdots \\ c_n(\tau_1, \tau_2, \cdots \tau_{n-1}) &\triangleq cum[x(k), x(k+\tau_1), x(k+\tau_2), \cdots, x(k+\tau_{n-1})] \end{aligned} \tag{2.14}$$

where *cum* represents cumulant. It is important to note that cumulants are closely related to moments (e.g., $c_2 = m_2 - m_1^2$) and can be calculated from the knowledge of moments. The computation of cumulants of order n requires all moments up to order n.

2.2.3 The relationship between moments and cumulants

Given a set of n random variables, $\{x_1, x_2, \ldots x_n\}$, the general relationship between their joint moments and joint cumulants of order n is given by:

$$cum[x_1, x_2, \ldots, x_n] = \sum_{p=1}^{n} (-1)^{p-1}\ (p-1)!\ E\{\prod_{i\in s_1} x_i\}\ E\{\prod_{i\in s_2} x_i\} \ldots\ E\{\prod_{i\in s_p} x_i\}, \tag{2.15}$$

where the summation extends over all partitions $(s_1, s_2, \ldots, s_p)$; $p = 1, 2, \ldots, n$, of the set of integers $(1, 2, \ldots, n)$ (Rosenblatt and Van Ness, 1965). For example, for the third order cumulant, the set of integers (1, 2, 3) can be partitioned into

$p = 1$	$s_1 = \{1, 2, 3\}$		
$p = 2,$	$s_1 = \{1\}$	$s_2 = \{2, 3\}$	
	$s_1 = \{2\}$	$s_2 = \{1, 3\}$	
	$s_1 = \{3\}$	$s_2 = \{1, 2\}$	
$p = 3,$	$s_1 = \{1\}$	$s_2 = \{2\}$	$s_3 = \{3\}$

Therefore, using Equation (2.15) we can write,

$$\begin{aligned} cum[x_1, x_2, x_3] &= E\{x_1x_2x_3\} - E\{x_1\}E\{x_2x_3\} - E\{x_2\}E\{x_1x_3\} \ldots \\ &\quad -E\{x_3\}E\{x_1x_2\} + 2E\{x_1\}E\{x_2\}E\{x_3\} \end{aligned} \tag{2.16}$$

If $x_2 = x_3 = x_1 = x$, then

$$\begin{aligned} cum[x, x, x] &= E\{x^3\} - 3E\{x\}E\{x^2\} + 2\{E\{x\}\}^3 \\ &= m_3 - 3m_1m_2 + 2m_1^3 \end{aligned} \tag{2.17}$$

which is the same as the expression for c_3 obtained earlier.

Note that the computation of joint cumulant of order r requires the knowledge of all moments up to order r.

2.2.4 Properties of moments and cumulants

1. $mom[a_1x_1, a_2x_2, \ldots, a_nx_n] = (a_1a_2\ldots,a_n)\ mom[x_1, x_2, \ldots, x_n]$, and $cum[a_1x_1, a_2x_2, \ldots, a_nx_n] = (a_1a_2 \ldots a_n)\ cum[x_1, x_2, \ldots, x_n]$, where $(a_1, a_2, \ldots, a_n)$ are constants, and *mom* represents moment.

2. Moments and cumulants are symmetric functions in their arguments, e.g., $mom[x_1, x_2, x_3] = mom[x_1, x_3, x_2] = mom[x_3, x_2, x_1]$, and so on.

3. If the random variables $(x_1, x_2, \ldots, x_n)$ can be divided into two or more statistically independent groups, then their n^{th} order cumulant is identically zero, i.e., $cum[x_1, x_2, \ldots, x_n] = 0$, whereas, in general, $mom[x_1, x_2, \ldots, x_n] \neq 0$.

4. If the sets of random variables $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are independent, then $cum[x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n] = cum[(x_1, x_2, \ldots, x_n)] + cum[y_1, y_2, \ldots, y_n]$, but in case of moments, $mom[x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n] \neq mom[(x_1, x_2, \ldots, x_n)] + mom[y_1, y_2, \ldots, y_n]$

5. If the set of random variables $(x_1, x_2, \ldots, x_n)$ is jointly Gaussian, then all information about their distribution is contained in the moments of order $n \leq 2$. Therefore, all moments of order greater than two have no additional information. This leads to the fact that all joint cumulants of order greater than two are identical to zero for Gaussian random vectors. Hence, cumulants of order greater than two measure the non-Gaussian nature of a time series.

Properties 4 and 5 demonstrate the ability of the cumulants to suppress noise in signal processing when the noise is additive Gaussian.

Example 1. Consider the random variables, $z_i = y_i + x_i$, for $i = 1, 2, 3$. If the joint probability density function of (y_1, y_2, y_3) is non-Gaussian and (x_1, x_2, x_3) is jointly

Gaussian and independent from (y_1, y_2, y_3) and also $E\{y_i\} \neq 0$, $E\{x_i\} \neq 0$, then using the fourth and fifth properties of cumulants, we get

$$cum[(z_1, z_2, z_3)] = cum[(y_1, y_2, y_3)] \tag{2.18}$$

since $cum[(x_1, x_2, x_3)] = 0$. In contrast, this is not true for moments. Therefore, cumulants of order greater than two can be conveniently used for analyzing signals that contain additive white noise.

Example 2. Consider the random variable, $y = x + e$, where y is a measured signal, x is the true signal, and e is the noise and is independent of x.

Moments: First order moment of y,

$$m_1 = E[y] = E[x] + E[e] \tag{2.19}$$

Second order moment of y,

$$m_2 = E[y^2] = E[(x+e)^2] = E[x^2] + 2E[x]\,E[e] + E[e^2] \tag{2.20}$$

Third order moment of y,

$$\begin{aligned} m_3 &= E\{y^3\} \\ &= E\{(x+e)^3\} \\ &= E\{x^3\} + 3E\{x^2\}E\{e\} + 3E\{x\}E\{e^2\} + E\{e^3\} \\ &= mom[x,x,x] + 3mom[x,x]\ mom[e] \\ &\quad +3\ mom[x]\ mom[e,e] + mom[e,e,e] \end{aligned} \tag{2.21}$$

Cumulants: First order cumulant of y,

$$cum[y] = E\{y\} = E[x] + E[e] \tag{2.22}$$

Second order cumulant of y,

$$\begin{aligned} cum[y,\ y] &= E[y^2] - \{E[y]\}^2 \\ &= E[(x+e)^2] - \{E[x] + E[e]\}^2 \\ &= E[x^2] + 2E[x]\,E[e] + E[e^2] - \{E[x]\}^2 - 2E[x]\,E[e] - \{E[e]\}^2 \\ &= E[x^2] + E[e^2] - \{E[x]\}^2 - \{E[e]\}^2 \end{aligned} \tag{2.23}$$

Third order cumulant of y,

$$\begin{aligned} cum[y,\, y,\, y] &= E\{y^3\} - 3E\{y\}E\{y^2\} + 2[E\{y\}]^3 \\ &= E\{(x+e)^3\} - 3E\{(x+e)\}E\{(x+e)^2\} + 2[E\{(x+e\}]^3 \\ &= E\{x^3\} - 3E\{x^2\}E\{x\} + 2[E\{x\}]^3 \\ &\quad + E\{e^3\} - 3E\{e^2\}E\{e\} + 2[E\{e\}]^3 \\ &= cum[x,x,x] + cum[e,e,e] \end{aligned} \tag{2.24}$$

If the noise is zero-mean and Gaussian distributed with the following properties:

$$\begin{aligned} \text{the first moment of e:} \quad & E\{e\} = 0 \\ \text{the second moment of e:} \quad & E\{e^2\} = \sigma_e^2 \\ \text{the third moment of e:} \quad & E\{e^3\} = 0 \end{aligned} \tag{2.25}$$

First order moment of y,

$$m_1 = E[y] = E[x] \tag{2.26}$$

Second order moment of y,

$$m_2 = E[y^2] = E[x^2] + \sigma_e^2 \tag{2.27}$$

Third order moment of y,

$$m_3 = mom[x,x,x] + 3\, E[x]\, \sigma_e^2 \tag{2.28}$$

First order cumulant of y is

$$cum[y] = E\{y\} = E[x] \tag{2.29}$$

Second order cumulant of y,

$$\begin{aligned} cum[y,\, y] &= E[y^2] - \{E[y]\}^2 \\ &= E[x^2] + \sigma_e^2 - \{E[x]\}^2 \end{aligned} \tag{2.30}$$

Third order cumulant of y,

$$cum[y,\, y,\, y] = cum[x,x,x] \tag{2.31}$$

From the above it is clear that cumulants of order greater than 2 are unaffected by the Gaussian noise, but the same order moments are not.

2.2.5 Moments and cumulants of stationary signals

If $x(k)$, $k = \pm 1, \pm 2, \ldots,$ is a real stationary random signal and its moments up to order n exist, then the n^{th} order moment of x is defined as

$$m_n(\tau_1, \tau_2, \cdots \tau_{n-1}) \triangleq E[x(k)x(k+\tau_1)x(k+\tau_2)\cdots x(k+\tau_{n-1})] \tag{2.32}$$

where, $\tau_i = 0, \pm 1, \pm 2, \ldots$.
Similarly, the n^{th} order cumulant of x is defined as

$$c_n(\tau_1, \tau_2, \cdots \tau_{n-1}) \triangleq cum[x(k), x(k+\tau_1), x(k+\tau_2), \cdots, x(k+\tau_{n-1})] \tag{2.33}$$

Combining Equations 2.33, 2.32, and 2.15, the following relationships of moment and cumulant sequences of $x(k)$ can be obtained:
1^{st} Order cumulant:

$$c_1 = m_1 = E[x(k)] \tag{2.34}$$

2^{nd} Order cumulant:

$$c_2(\tau_1) = m_2(\tau_1) - m_1^2 = m_2(-\tau_1) - m_1^2 = c_2(-\tau_1) \tag{2.35}$$

3^{rd} Order cumulant: Combining Equations 2.33, 2.32, and 2.16

$$c_3(\tau_1, \tau_2) = m_3(\tau_1, \tau_2) - m_1[m_2(\tau_1) + m_2(\tau_2) + m_2(\tau_2 - \tau_1)] + 2m_1^3 \tag{2.36}$$

Similarly, the other higher order cumulants can be obtained.

2.3 Spectral Analysis

Not all the information content of a signal can necessarily be obtained easily from time domain analysis of the data. Transforming the signal from time to frequency domain can expose the periodicities of the signal, detect nonlinearities present in the signal and aid in understanding the signal generating process. The Discrete Fourier Transform (DFT), which plays a key role in modern digital signal processing, is the main tool for this type of transformation.

For a real strictly stationary signal, $\{x(k)\}$, $k = 0, \pm 1, \pm 2, \ldots$ the n^{th} order cumulant sequence is defined by

$$c_n(\tau_1, \tau_2, \ldots, \tau_{n-1}) = cum[x(k),\ x(k+\tau_1),\ \ldots,\ x(k+\tau_{n-1})] \tag{2.37}$$

Assuming that the cumulant sequence satisfies the condition $\sum_{\tau_1=-\infty}^{\infty} \cdots \sum_{\tau_{n-1}=-\infty}^{\infty} (1+|\tau_i|)|c_n(\tau_1,\tau_2,\ldots,\tau_{n-1})| < \infty$, for i= 1, 2, ..., n-1, the n^{th} order cumulant spectrum $c_n(f_1,f_2,\ldots,f_{n-1})$ of $\{x(k)\}$ is defined as the $(n-1)$ dimensional Fourier Transform of the n^{th} order cumulant sequence, i.e.,

$$c_n(f_1,f_2,\ldots,f_{n-1}) = \sum_{\tau_1=-\infty}^{\infty} \cdots \sum_{\tau_{n-1}=-\infty}^{\infty} c_n(\tau_1,\tau_2,\ldots,\tau_{n-1}) \exp\{-j(f_1\tau_1+f_2\tau_2+\ldots+f_{n-1}\tau_{n-1})\} \tag{2.38}$$

where $|f_i| \leq \pi$, for $i = 1,2,\ldots,n-1$ and $|f_1+f_2+\ldots+f_n| \leq \pi$

The power spectrum, bispectrum and trispectrum are special cases of the n^{th} order cumulant spectrum.

2.3.1 Power spectrum, $n = 2$

The power spectrum is the frequency domain counterpart of the second order moment or cumulant of a signal, and is usually obtained from a zero-mean time series. For such a time series, say $x(k)$, the second order moment and the second order cumulant are identical. The power spectrum can be obtained using either of the following two methods:

1. **Indirect Method:** First, the power spectrum can be calculated from the DFT of the second order moments or cumulants of the time series. This is known as the indirect method. This method involves the following two steps:

   (a) Calculate the second order moment sequence using the following equation:

   $$m_2(\tau) \triangleq E[x(k)x(k+\tau)] \equiv c_2(\tau) \tag{2.39}$$

   (b) If m_2 is absolutely summable, i.e., the summation of the absolute values of m_2 at all lags is a finite number, then the power spectrum is given by:

   $$P(f) = DFT[m_2(\tau)] \equiv \sum_{\tau=0}^{N-1} m_2(\tau)e^{-j2\pi\tau f/N} \tag{2.40}$$

2. **Direct Method:** The power spectrum can also be calculated from the direct Fourier transformation of the time series. This is known as the direct method. It also involves two steps:

   (a) Perform the DFT of the data series.

   $$X(f) = \sum_{k=0}^{N-1} x(k)e^{-j2\pi kf/N} \tag{2.41}$$

(b) Next calculate the power spectrum as

$$P(f) = E[X(f)X(-f)] \equiv E[X(f)X^*(f)] \equiv E[|X(f)|^2] \tag{2.42}$$

where the * denotes a complex conjugate.

The power spectrum can be thought of as a decomposition or spread of the signal energy over the frequency channels obtained from the length of the Fast Fourier Transform. From the expressions used to estimate the power spectrum, it is clear that the power spectrum is phase blind, i.e., the phase information is ignored. The information obtained from the power spectrum or auto-correlation sequence, or more generally from/up to the second order statistics are sufficient only for complete description of linear processes. In order to study the deviations from Gaussianity and the presence of nonlinearities in signal generating processes, it is necessary to look beyond the second order statistics (SOS). This brings us to the domain of Higher Order Statistics (HOS). All statistics of order greater than 2 are called HOS.

2.3.2 Bispectrum, $n = 3$

The methodology used for estimating the second order spectrum or power spectrum can easily be extended to obtaining frequency domain counterparts of higher order cumulants. For example, the bispectrum is the frequency domain representation of the third order cumulant. It is defined as

$$B(f_1, f_2) \triangleq DDFT[c_3(\tau_1, \tau_2)] \equiv E[X(f_1)X(f_2)X^*(f_1 + f_2)] \tag{2.43}$$

where DDFT stands for Double Discrete Fourier Transformation.
Similary, the trispectrum ($n = 4$) is defined as

$$T(f_1, f_2, f_3) \triangleq TDFT[c_4(\tau_1, \tau_2, \tau_3)] \equiv E[X(f_1)X(f_2)X(f_3)X^*(f_1 + f_2 + f_3)] \tag{2.44}$$

where TDFT stands for Triple Discrete Fourier Transformation. These higher order spectra are known as polyspectra. This thesis involves the extensive use of the bispectrum. Therefore, for a thorough understanding of this term, the next section is devoted toward its interpretation.

Equation 2.43 shows that the bispectrum is a complex quantity having both magnitude and phase. It can be plotted against two independent frequency variables, f_1 and f_2, in a three dimensional figure. Just as the discrete power spectrum has a point of symmetry

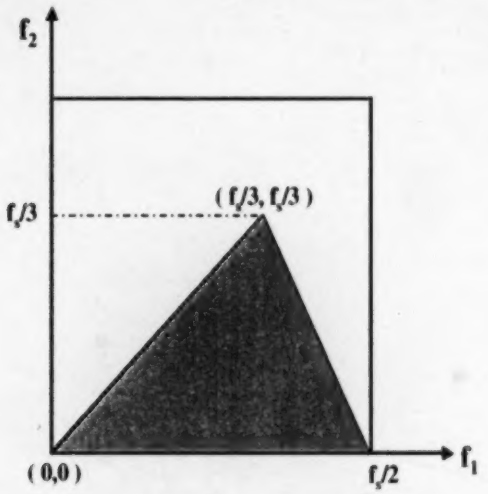


Figure 2.1: *Principal domain of the bispectrum.*

at the folding frequency, the discrete bispectrum also has 12 planes of symmetries in the (f_1,f_2) plane. For details, readers are referred to (Rosenblatt and Van Ness, 1965; Nikias and Raghuveer, 1987; Kim and Powers, 1979; Nikias and Petropulu, 1993). Examination of the bispectrum in only one region gives sufficient information. The other regions of the (f_1,f_2) plane are redundant. The shaded triangular region in Figure 2.1 shows the non-redundant principal domain of the bispectrum. Throughout this work frequencies are normalized such that $f_s = 1$, where f_s is the sampling frequency. Therefore, the Nyquist frequency is $f_N = 0.5$. Each point in the plot represents the bispectral content of the signal at the bifrequency, (f_1,f_2). In fact, the bispectrum at point $(B(f_1,f_2),f_1,f_2)$ measures the interaction between frequencies f_1 and f_2. This interaction between frequencies can be related to non-linearities present in the signal generating systems (Fackrell, 1996), and therein lies the core of its usefulness in the detection and diagnosis of non-linearities.

2.3.3 Estimation of the bispectrum

In practice, the higher order spectra of a signal have to be estimated from a finite set of measurements. The underlying methods for polyspectra estimation are simply extensions of the well-established power spectrum estimation methods. Essentially, there are two broad non-parametric approaches: (1) the indirect method, based on estimating

the cumulant functions and then taking the Fourier Transform; and (2) the direct method, based on a segment averaging approach. The details of these two approaches are discussed below:

1. **Indirect method:** In this method the data series are segmented into K records and the third order cumulants are estimated for each of the records. The average of the third order cumulants are taken over the K segments. The bispectrum is then obtained by taking the Fourier transformation of the average third order cumulants together with the appropriate data window function. The steps are as follows:

   (a) The data series $x(k)$, $k = 0,1,...,N-1$ is divided into K segments, $i = 0,1,...,K-1$ each of length M. These segments can overlap, so $K \geq N/M$. Let the i^{th} segment of $x(k)$ be $x_i(k)$, $k = 0,1,...,M-1$.

   (b) The mean μ_i of the i^{th} segment is calculated and subtracted from each sample in the segment.

$$\mu_i = \frac{1}{M}\sum_{k=0}^{M-1} x_i(k) \tag{2.45}$$

$$x_i'(k) = x_i(k) - \mu_i \tag{2.46}$$

   (c) Obtain an estimate of the third order cumulants for each data segment

$$c_{3,i}(m,n) = \frac{1}{M}\sum_{k=s_1}^{s_2} x_i'(k)x_i'(k+m)x_i'(k+n) \tag{2.47}$$

   where,

   $s_1 = max(0,-m,-n)$

   $s_2 = min(M-1, M-1-m, M-1-n)$

   (d) Average $c_{3,i}(m,n)$ over all segments

$$\bar{c}_3(m,n) = \frac{1}{K}\sum_{i=1}^{K} c_{3,i}(m,n) \tag{2.48}$$

   (e) Obtain the bispectrum estimate

$$\hat{B}(f_1,f_2) = \sum_{m=-L}^{L}\sum_{n=-L}^{L} \bar{c}_3(m,n)\, w(m,n)\, e^{\{-j(mf_1+nf_2)\}} \tag{2.49}$$

   where $L < M-1$ and $w(m,n)$ is a two-dimensional window function. Window functions should possess the same symmetry properties as the third order

cumulants. Detailed descriptions of some window functions are available in the literature (Nikias and Petropulu, 1993; Nikias and Raghuveer, 1987; Mendel, 1991).

2. **Direct method:** The direct method of bispectrum estimation is an extension of the Welch periodogram averaging technique for spectral estimation (Fackrell, 1996). The method is depicted schematically in Figure 2.2 and consists of the following steps:

   (a) The data series $x(k)$, $k = 0, 1, ..., N-1$ is divided into K segments, $i = 0, 1, ..., K-1$ each of length M. These segments can overlap, so that $K \geq N/M$. Let the i^{th} segment of $x(k)$ be $x_i(k)$, $k = 0, 1, ..., M-1$.

   (b) The mean μ_i of the i^{th} segment is calculated and subtracted from each sample in the segment.

$$\mu_i = \frac{1}{M} \sum_{k=0}^{M-1} x_i(k) \tag{2.50}$$

$$x_i'(k) = x_i(k) - \mu_i \tag{2.51}$$

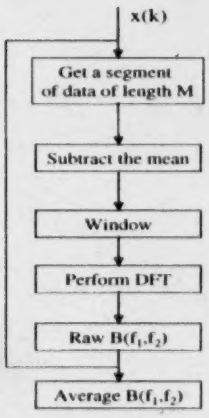


Figure 2.2: *Information flow diagram for the direct method of bispectrum estimation.*

   (c) The zero mean-centered segment of the data x_i' is then multiplied by a suitable data window $w(k)$, which provides some control over the spectral leakage. This window function may be a boxcar, Hamming, Hanning or any other window used in ordinary spectral estimation.

$$x_i''(k) = w(k)x_i'(k) \tag{2.52}$$

(d) For each segment, compute the DFT, $X_i(f)$

$$X_i(f) = \sum_{k=0}^{M-1} x_i''(k) e^{-j2\pi kf/M} \tag{2.53}$$

where f is the discrete frequency. From this DFT the raw spectral estimates of $P_i(f)$ and the bispectral estimates of $B_i(f_1, f_2)$ can be found.

$$\hat{P}_i(f) = X_i(f) X_i^*(f) \tag{2.54}$$

$$\hat{B}_i(f_1, f_2) = X_i(f_1) X_i(f_2) X_i^*(f_1 + f_2) \tag{2.55}$$

(e) Then the raw estimates from all K segments can be averaged to give the following estimates

$$\hat{P}(f) = \frac{1}{K} \sum_{i=0}^{K-1} \hat{P}_i(f) \tag{2.56}$$

$$\hat{B}(f_1, f_2) = \frac{1}{K} \sum_{i=0}^{K-1} \hat{B}_i(f_1, f_2) \tag{2.57}$$

2.3.4 Properties of estimators and asymptotic behaviour

The statistical properties of indirect and direct methods of higher order spectral estimation have been studied extensively (Rosenblatt and Van Ness, 1965; Brillinger and Rosenblatt, 1967*a*; Brillinger and Rosenblatt, 1967*b*; Rao and Gabr, 1984; Nikias and Petropulu, 1993; Riggs, 1999). Assume that $P(f)$ and $B(f_1, f_2)$ are the true power spectrum and bispectrum, respectively, of a stationary zero-mean signal, $x(k)$. Let $\hat{B}(f_1, f_2)$ be the estimate of $B(f_1, f_2)$ using either the indirect method or the direct method and a single realization of the signal, $x(k)$ of length N. It has been shown mathematically (Nikias and Mendel, 1993; Kim and Powers, 1979) that for sufficiently large record size, M, and total length, N, both the direct and indirect methods provide approximately unbiased estimates, namely:

$$E\{\hat{B}(f_1, f_2)\} \cong B(f_1, f_2) \tag{2.58}$$

with asymptotic variances

$$var\{\Re[\hat{B}(f_1, f_2)]\} \cong var\{\Im[\hat{B}(f_1, f_2)]\} \cong \frac{1}{2}\sigma^2(f_1, f_2) \tag{2.59}$$

where,

$$\sigma^2(f_1, f_2) = \begin{cases} \frac{VL^2}{MK} P(f_1) P(f_2) P(f_1 + f_2) & \text{(indirect)} \\ \\ \frac{N_0^2}{MK} P(f_1) P(f_2) P(f_1 + f_2) & \text{(direct)} \end{cases} \tag{2.60}$$

where K is the number of records, M is the number of samples per record, V is the total energy of the window used for bispectrum estimation (for a rectangular window V is unity), L is defined in step 5 of the indirect method. Because we did not use any frequency domain smoothing, in our case N_0 is same as M. From the above equations it is apparent that if a rectangular window is used and if $L = N_0$, both methods provide approximately the same estimates.

Brillinger and Rosenblatt (1967*a*) showed that when M and N are large, the error bicoherence is approximately complex Gaussian with zero mean and unity variance:

$$\frac{\hat{B}(f_1, f_2) - B(f_1, f_2)}{\sigma^2(f_1, f_2)} \sim N_c(0, 1) \tag{2.61}$$

where N_c denotes a complex normal distribution, i.e., both real and imaginary parts are normally distributed. Another important conclusion that follows from the asymptotic results developed by (Brillinger and Rosenblatt, 1967*a*; Brillinger and Rosenblatt, 1967*b*) is that these statistics can be treated as independent random variables over the grid of principal domain if the grid width is larger than or equal to the bispectrum bandwidth; i.e., $\hat{B}(f_q, f_p)$ and $\hat{B}(f_r, f_s)$ are independent for $q \neq r$ or $p \neq s$ if $|f_{q+1} - f_q| \geq \triangle_0$ or $|f_{r+1} - f_r| \geq \triangle_0$, where,

$$\triangle_0 = \begin{cases} \frac{1}{L} & \text{(indirect)} \\ \\ \frac{1}{N_0} & \text{(direct)} \end{cases} \tag{2.62}$$

2.3.5 Bicoherence or normalized bispectrum

As shown in Equation 2.60, the bispectral estimates are asymptotically unbiased and the variance of the estimator depends on the second order spectral properties (Hinich, 1982). That is,

$$var(\hat{B}(f_1, f_2)) \propto P(f_1)P(f_2)P(f_1 + f_2) \tag{2.63}$$

Since the estimate depends directly on signal energies in the bifrequency, the variance of the estimate will be higher at a bifrequency where the energy is high and lower where the energy is low. This causes a serious problem in the estimation. This unsatisfactory property can be resolved in several ways. One way is to prewhiten the signal prior to bispectral analysis (Collis *et al.*, 1998). However, an easier solution is to normalize the bispectrum to get a new measure whose variance is independent of the signal energy. Hinich (1982) has suggested normalization of the bispectrum using the following expression called the skewness function.

$$s^2(f_1, f_2) \triangleq \frac{|E[B(f_1, f_2)]|^2}{E[P(f_1)]E[P(f_2)]E[P(f_1 + f_2)]} \tag{2.64}$$

where $s^2(f_1, f_2)$ is known as the skewness function. The major drawback of this normalization is that the magnitudes of this function are not bounded. The only reason for dividing the bispectrum with power spectrums is to remove the undesirable variance properties of the estimator. This method of normalization has been extensively used by Hinich (1982) for statistical tests of Gaussianity and linearity of a signal. The Higher Order Spectral Analysis (HOSA) toolbox in MATLAB® (Swami *et al.*, 1993) has also adopted this normalization. This normalized bispectrum is defined as bicoherence in this toolbox. However, it would be more appropriate to call it "skewness function".

The bicoherence is better defined by the following equation :

$$bic^2(f_1, f_2) \triangleq \frac{|B(f_1, f_2)|^2}{E[|X(f_1)X(f_2)|^2]\ E[|X(f_1 + f_2)|^2]} \tag{2.65}$$

where *'bic'* is the bicoherence function. Kim and Powers (1979) have shown that the variance of the bicoherence estimator roughly satisfies the following expression:

$$var[\hat{bic}^2(f_1, f_2)] \approx \frac{1}{M}[1 - bic^2(f_1, f_2)] \tag{2.66}$$

where M is defined as the number of segments used in the estimation. Note that it is a consistent estimator in the sense that the variance approaches zero as M approaches infinity.

A useful feature of the bicoherence function is that it is bounded between 0 and 1. This can be demonstrated using the Cauchy-Schwartz inequality which may be expressed as,

$$|E[z_1, z_2]|^2 \leq E[|z_1|^2]E[|z_2|^2] \tag{2.67}$$

Choosing $z_1 = X(f_1)X(f_2)$ and $z_2 = X(f_1 + f_2)$, it can be shown that

$$0 \leq bic^2 \leq 1 \tag{2.68}$$

There are also other normalization methods (Fackrell, 1996) but are not widely used because their properties have not been extensively studied.

Estimation of the Squared Bicoherence

To estimate the squared bicoherence from a finite set of measurements, the method of bispectrum estimation can be followed directly. Because we have used the direct method throughout this work, only the direct method of estimation of bicoherence is discussed here.

1. Using the first four steps of the direct method of bispectrum estimation, obtain the estimates of $\hat{P}_i(f) = X_i(f)X_i^*(f)$ and $\hat{B}_i(f_1, f_2)$ for each segment of the data.

2. Equation 2.65 can be rewritten as

$$bic^2(f_1,f_2) \triangleq \frac{|E[X(f_1)X(f_2)X^*(f_1+f_2)]|^2}{E[|X(f_1)X(f_2)|^2]E[|X(f_1+f_2)|^2]} \tag{2.69}$$

Now, the expectation operator can be replaced with a summation operator over the number of data segments in the following way:

$$\hat{bic}^2(f_1,f_2) = \frac{|\frac{1}{M}\sum_{i=1}^{M} X_i(f_1)X_i(f_2)X_i^*(f_1+f_2)|^2}{\frac{1}{M}\sum_{i=1}^{M}|X_i(f_1)X_i(f_2)|^2 \;\; \frac{1}{M}\sum_{i=1}^{M}|X_i(f_1+f_2)|^2} \tag{2.70}$$

The squared bicoherence can be estimated using the above equation.

2.3.6 Properties of bispectrum and bicoherence

1. *The theoretical bispectrum of a Gaussian signal is zero*

The moment generating function for a Gaussian signal $x(t)$ is given by (Stuart and Ord, 1987):

$$M_x(t) = exp[\mu t + \frac{1}{2}\sigma^2 t^2] \tag{2.71}$$

Therefore, the cumulant generating function for x is

$$C_x(t) = \ln M_x(t) = \mu t + \frac{1}{2}\sigma^2 t^2 \tag{2.72}$$

Now, Equation 2.9 can be rewritten as

$$C_x(t) = c_1 t + \frac{1}{2!}c_2 t^2 + \frac{1}{3!}c_3 t^3 + \ldots + \frac{1}{r!}c_r t^r + \ldots \tag{2.73}$$

comparing Equations 2.72 and 2.73, we obtain

$$\begin{aligned} c_1 &= \mu \\ c_2 &= \sigma^2 \\ c_r &= 0, \quad r > 2 \end{aligned} \tag{2.74}$$

Thus we see that for any Gaussian signal the cumulants at the zero$^t h$ lag of order greater than 2 (e.g., $c_3(0,0)$, $c_4(0,0,0)$, and higher) are identically zero. This result can be easily generalized to other non-zero lags cumulants. As for any auto-covariance sequence, $c_2(0) > c_2(\tau)$, $\tau \neq 0$, the same is true for other cumulants of higher order. For example, $c_2(0,0) > c_2(\tau_1,\tau_2)$, $\tau_1 = \tau_2 \neq 0$. If $c_2(0,0) = 0$, then $c_2(\tau_1,\tau_2)$, $\tau_1 = \tau_2 \neq 0$ must be equal to zero. The same is true for other higher order cumulants.

The bispectrum is the frequency domain counterpart of the third order cumulants. Because third order cumulants for a Gaussian signal are identically zero, the bispectrum will also be zero for a Gaussian signal.

2. *The theoretical bicoherence of a Gaussian signal is identically zero.*

   This follows directly from the property 1. Because the bicoherence or the skewness function is a scaled bispectrum, a zero bispectrum leads to a zero bicoherence or skewness.

3. *The theoretical bispectrum of a non-Gaussian signal is blind to the additive Gaussian noise.*

   It was shown in Section 2.2.4 that the third order cumulants of a non-Gaussian signal is independent of the additive Gaussian noise. This property follows directly from it.

4. *The theoretical bicoherence of a signal is not blind to Gaussian noise.*

   Because the bicoherence or skewness function is scaled with the power of the signal, which is not independent of noise, the bicoherence is affected by the additive Gaussian noise.

5. *If a signal is filtered by a linear filter, then the magnitude of the bicoherence is unchanged.*

   Consider a signal, $x(n)$, filtered with a linear causal time-invariant filter, $h(k)$, to obtain the signal, $y(n)$. This can be represented by the following equation:

$$y(n) = \sum_{k=0}^{\infty} h(k)x(n-k) \tag{2.75}$$

   In the frequency domain, this can be rewritten as

$$Y(f) = H(f)X(f) \tag{2.76}$$

   Now, substituting this in the defining equation of bicoherence (Equation 2.65), we

obtain

$$
\begin{aligned}
bic^2(f_1,f_2) &\triangleq \frac{|E[Y(f_1)Y(f_2)Y^*(f_1+f_2)]|^2}{E[|Y(f_1)Y(f_2)|^2]E[|Y(f_1+f_2)|^2]} \\
&= \frac{|E[H(f_1)X(f_1)H(f_2)X(f_2)H^*(f_1+f_2)X^*(f_1+f_2)]|^2}{E[|H(f_1)X(f_1)H(f_2)X(f_2)|^2]E[|H(f_1+f_2)X(f_1+f_2)|^2]} \\
&= \frac{|H(f_1)H(f_2)H^*(f_1+f_2)|^2}{|H(f_1)H(f_2)|^2\,|H(f_1+f_2)|^2} \times \\
&\quad \frac{|E[X(f_1)X(f_2)X^*(f_1+f_2)]|^2}{E[|X(f_1)X(f_2)|^2]E[|X(f_1+f_2)|^2]} \\
&= \frac{|E[X(f_1)X(f_2)X^*(f_1+f_2)]|^2}{E[|X(f_1)X(f_2)|^2]E[|X(f_1+f_2)|^2]} \qquad (2.77)
\end{aligned}
$$

(since H is time invariant, it can be taken out of expectation operator and all H's cancel out)

Note that the above only holds when $H(f) > 0,\ \forall\ f$. Otherwise, 0/0 may occur in the calculation of above quantity. This is equivalent to the requirement that the linear filter should not have any zeros on the unit circle [(Fackrell, 1996)].

This property enables us to use bicoherence to detect the linearity of a signal generating process.

6. *The theoretical bicoherence of a harmonic signal that can be decomposed into a sum of sinusoids shows peaks if the signal possesses the property of 'Quadratic Phase Coupling (QPC)'.*

   A signal is said to have QPC if the signal phases, for example, ϕ_1, ϕ_2, and ϕ_3 at frequencies f_1, f_2, $f_3 = f_1 + f_2$ respectively, have the relation $\phi_1 + \phi_2 = \phi_3$. This is an indicator of a nonlinear signal generating mechanism. A system which possesses quadratic nonlinearity, for example a square function, can yield a QPC signal. This has been illustrated with two examples in Section 3.4.

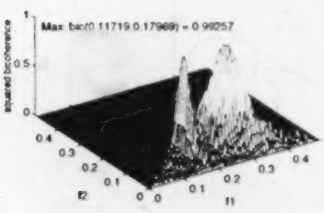
2.4 Bispectrum or Bicoherence Estimation Issues

The main estimation parameters that need to be chosen for bispectral analysis are the same parameters required for second order spectral analysis. Examples include the choice of window function, data length, data segment length, length of fourier transform, and overlapping or non-overlapping windows (Fackrell, 1996). In this section, some of these issues will be addressed and discussed using the following illustrative example.

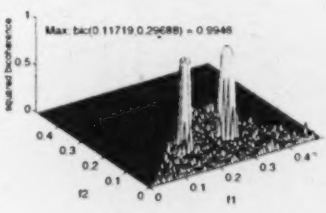
Illustrative Example. This example consists of the following signal.

$$y(t) = sin(2\pi f_1 t + \phi_1) + sin(2\pi f_2 t + \phi_2) + sin(2\pi f_3 t + \phi_3) + sin(2\pi f_4 t + \phi_4) + n(t) \quad (2.78)$$

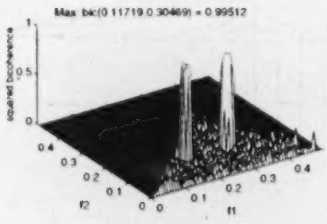
where the values of f_1, f_2, f_3, and f_4 are 0.12, 0.18. 0.30 and 0.42 respectively; the values of ϕ_1, ϕ_2, ϕ_3, ϕ_4 are $\frac{\pi}{3}$, $\frac{\pi}{12}$, $\frac{\pi}{4}$, $\frac{3\pi}{8}$ respectively; $n(t)$ is a zero-mean white noise signal with variance 0.2, and t is time from 1 to 4096 s. The signal $y(t)$ is a quadratic phase coupled signal because its frequencies have the relations $f_1 + f_2 = f_3$ and $f_1 + f_3 = f_4$, and its phases have the relations $\phi_1 + \phi_2 = \phi_3$ and $\phi_1 + \phi_3 = \phi_4$. Therefore, these two phase couplings at bifrequencies (0.12 , 0.18) and (0.12 , 0.30) should appear in the bicoherence plot as two peaks at the same frequencies.



(a) Bicoherence using Box Car window



(b) Bicoherence using Hanning window



(c) Bicoherence using Hamming window

Figure 2.3: *Bicoherence plot for the QPC example.*

2.4.1 Choice of window function

In spectral analysis, the use of window function is very common. The main reason behind using a window is to solve the problem of spectral leakage that occurs between neighboring frequency channels of a peak. Spectral leakage is the term used to describe the loss of power of a given frequency to other frequency bins in the DFT. Spectral leakage can be visualized from the spread of the frequency components. Each frequency component of a signal should contribute only to one single frequency of the Fourier Transform called an FFT bin, but spectral leakage causes the energy to be spread to the neighboring frequencies. The window function controls the spreading. The contribution from any real frequency component to a given FFT bin is weighted by the amplitude of the window function's frequency spectrum centered at the FFT bin. Theories and issues related to the use of various types of windows in bispectral estimation have been addressed in (Fackrell, 1996; Nikias and Petropulu, 1993; Chandaran and Elgar, 1991). The performances of three window functions have been compared in (Fackrell, 1996) and it has been shown that the Hamming window was the most successful among them to best resolve the peak. The Hanning window stands next to the Hamming window in terms of peak resolution. The Box Car window function is the worst for resolving the peaks. Figure 2.3 shows the bicoherence plot for the time series of Equation 2.78. The figure shows that the Box Car function is the worst among the three windows in resolving peaks. This is because the kernel of the rectangular window has approximately two times wider sidelobes than that of the Hamming or Hanning window (Smith, 1998; Ifeachor and Jervis, 1993). A careful observation shows that the peaks in Figure 2.3(c) are better resolved than those in Figure 2.3(b).

2.4.2 Choice of data length, segment length, and fourier transform length

It is well known that bispectral estimates generally have higher variance than power spectral estimates for a given data length. The data length that is sufficient for a reliable power spectrum estimation may not be sufficient for a good bispectral estimation. Elgar and Guza (1988) reported empirical results for mean and variance of bicoherence estimates. Hinich (1982) suggested that if no frequency smoothing is used the data should be segmented in a way that the number of segments of data should be at least as large as the DFT size, i.e., $K \geq M$. In practical bispectral analysis, the length of data required depends

on signal to noise ratio. Fackrell (1996) presented a nice result for the dependency of theoretical bicoherence of a quadratically coupled sinusoids on the signal to noise ratio and DFT length. He stated the following: *If the effects of leakage are ignored, then the peak bicoherence $b^2(f_1, f_2)$ corresponds to the coupled components at frequencies f_1, f_2, and $f_1 + f_2$ of a quadratically coupled sinusoids consisting of three equal-amplitude coupled harmonics in variable levels of additive white Gaussian noise with a signal-to-noise ratio of SNR is given by:*

$$b^2(f_1, f_2) = \frac{1}{1 + \frac{18}{M} 10^{-SNR/10} + \frac{72}{M^2} 10^{-2SNR/10}} \tag{2.79}$$

where, *SNR* is defined as

$$SNR \triangleq 10 \log_{10} \frac{\sigma_x^2}{\sigma_n^2} \tag{2.80}$$

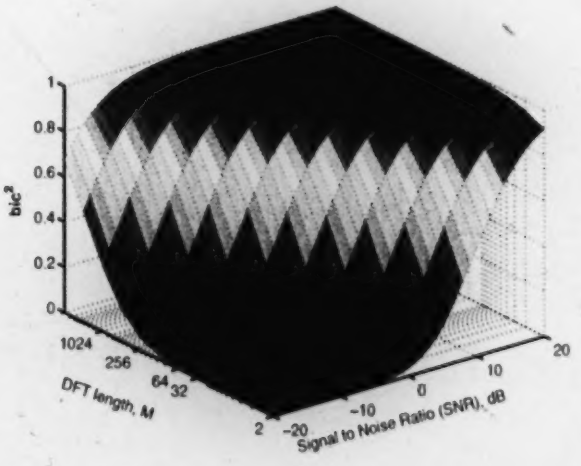


Figure 2.4: *The effect of signal to noise ratio and DFT length on bispectral estimation for a quadratically coupled harmonic signal. Larger DFT length and higher SNR provide better estimates.*

Equation 2.79 indicates how the bicoherence peak changes with *SNR* and segment length, *M*. If the *SNR* is very high, then the second and third terms in the denominator will be very close to zero and the bicoherence will be close to unity. As the *SNR* decreases, the bicoherence value also decreases. This can be presented in a 3-d plot as shown in Figure 2.4. The size of the peak of the squared bicoherence is dependent on *M* because, as

the DFT size increases, a better frequency resolution is obtained. Therefore, it is desirable to have the DFT size as large as possible. However, this requirement conflicts with the requirement of having a large number of data segments, K, to obtain reliable estimates. It also increases the computational load. The DFT length is usually chosen to be the same as each segment length. In order to increase the number of data segments for a better estimate, a certain amount of overlapping of data segments (e.g., 50% or less) can be allowed for data sets having short lengths.

2.5 Conclusions

This chapter has presented a tutorial introduction to Higher Order Statistics and its relation to classical second order statistics (SOS). Because the squared bicoherence function has played a central role in this thesis, a special emphasis has been given to the bispectrum and bicoherence, their estimation, and statistical properties of the estimator . Various properties have been explained with the help of illustrative examples wherever applicable.

3

Diagnosis of Poor Control Performance

The presence of nonlinearities in the control loop is one of the main reasons for poor performance of a linear controller designed based on linear control theory. The nonlinearity may be due to the presence of nonlinearities such as stiction, deadzone, hysteresis in the control valve and the nonlinear nature of the process itself. A nonlinear system often produces a non-Gaussian and nonlinear time series. The test of Gaussianity of a signal or the test of presence of nonlinearity in a system is a useful diagnostic aid towards determining the poor performance of a control loop. In this chapter two new indices, the Non-Gaussianity Index (*NGI*) and the Non-Linearity Index (*NLI*), based on HOS theory have been developed to detect and quantify signal non-Gaussianity and nonlinearity. These indices together with specific patterns in the process output (*pv*) vs. the controller output (*op*) plot can be conveniently used to diagnose the causes of poor control loop performance. The method has been successfully applied to many industrial data sets. Two of such case studies are presented here. In both cases, the results of the analysis were confirmed and mitigated during routine maintenance by plant engineers.

---

[1] *A preliminary version of this chapter was published in the proceedings of the 15th IFAC world congress, 2002 (Choudhury et al., 2002). A larger version including the diagnosis of poor control performance has been published in Automatica, Vol. 40, issue 10, 2004 (Choudhury et al., 2004d).*

3.1 Introduction

Increasing automation in modern process industries are taking place to achieve various objectives such as maintaining world class quality of the product, reducing operating and maintenance cost, enhancing operators' safety, meeting environmental and occupational health regulations, optimizing resource management, and increasing profitability from available resources. One manifestation of this increased automation is an increasing number of controllers and control loops in process industries. A process control plant may have anywhere from a few control loops to several thousands, depending on the complexity of the plant from the perspective of control. The performance assessment and monitoring of the performance of these control loops are crucial to the achievement of desired objectives. Performance demographics of 26,000 PID controllers collected over a period of two years and across a large range of continuous process industries have been discussed by (Desborough and Miller, 2002). The results from their paper are shown in Figure 3.1.

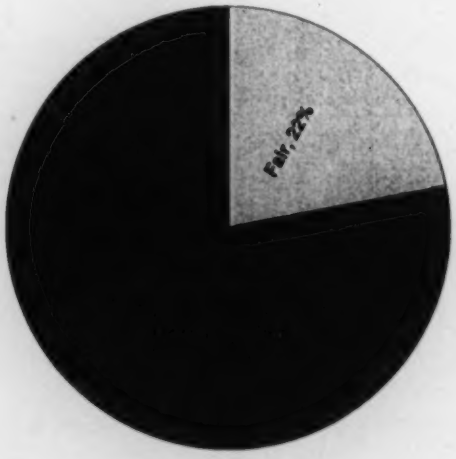


Figure 3.1: *Global multi-industry performance demographics.*

Each control loop (flow, pressure, level, temperature, etc.) was classified into one of the five categories — excellent, acceptable, fair, poor and open loop— based on a combined algorithm of minimum variance benchmark and oscillation metric. The classifications were further refined through extensive validation and feedback from industry to reflect controller performance relative to practical expectations for each measurement type. It is evident from Figure 3.1 that only a third of the loops are performing well or in an acceptable fashion.

The other two-thirds have significant opportunity for improvement. The key to improving their performance is to diagnose the causes behind their poor performance.

3.2 Problem Formulation

Figure 3.2 shows a typical control loop under feedback configuration. The objective of this control loop may be either set point tracking or disturbance rejection. In practice, data for three measurements per control loop are available. The available data are set point (*SP*), controlled variable (*PV*), and controller output (*OP* or sometimes also termed *OP*). There (Desborough and Harris, 1993; Harris, 1989; Stanfelj *et al.*, 1993) discussed various methods, for example, minimum variance benchmark and settling time benchmark, to estimate controller performance or loop performance from routine operating data. Also, this information is available from plant engineers or operators, who are dissatisfied with poorly performing loops. The challenge here is to identify the root cause of a poorly performing loop from routine operating data.

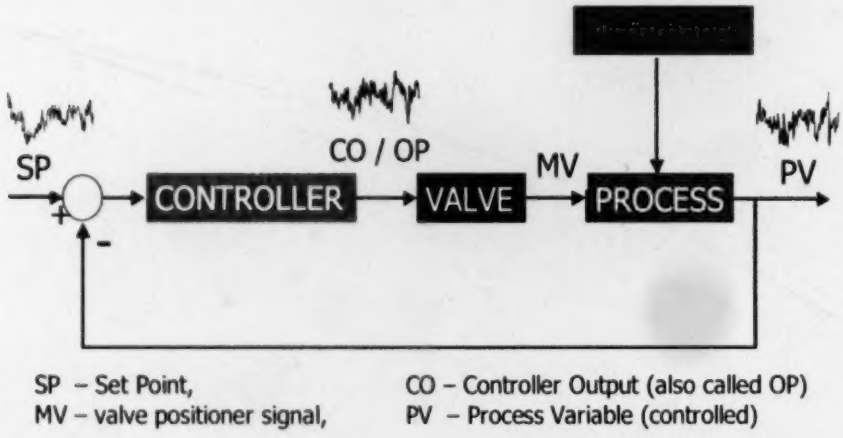


Figure 3.2: *A typical control loop under feedback control.*

Poor performance of a control loop is usually caused by poor controller tuning, presence of disturbances, and/or loop non-linearities. Loop nonlinearities cause poor performance because controllers are usually designed based on linear control theory assuming everything in the loop is locally linear. The nonlinearity in a loop may arise

due to the presence of valve nonlinearities such as stiction, deadzone, hysteresis in the control valve and/or nonlinear nature of the process itself. Such a nonlinear system often produces a non-Gaussian and nonlinear time series.

3.3 Test of Gaussianity and Linearity of a Signal

The test of the Gaussianity of a signal or for the presence of nonlinearity in a system can serve as a useful diagnostic tool to analyze poor performance of a control loop. Over the last few decades, many researches have used the notion of bispectrum to test Gaussianity and linearity of a time series (Hinich, 1982; Rao and Gabr, 1980; Collis *et al.*, 1998; Terdik and Máth, 1998; Yuan, 1999). Hinich (1982) and Yuan (1999) constructed statistical hypothesis test using sample interquartile range of the skewness function (for the definition of skewness function please refer to Section 2.3.5) for testing linearity of a time series. Rao and Gabr (1980) used bispectral density function (bispectrum) to construct a Hotelling T^2 test for the same purpose. The problem of using bispectrum or skewness function for testing linearity of time series is that the magnitude of this function is not bounded. It is difficult to compare magnitudes of an unbounded function when applied to multiple time series or signals. Terdik and Máth (1998) also used bispectrum to test the null hypothesis that a predictor of a time series is linear against the alternative that the predictor is quadratic. This is a narrow and less powerful test because often times the type of nonlinearity of a signal is unknown. Collis *et al.* (1998) used the bicoherence function whose magnitude is bounded between 0 and 1 to check linearity of a signal or time series. But they have not constructed any statistical test. In this thesis, bicoherence has been used to construct a chi-square test for examining linearity of a time series or signal. The hypothesis tests are incorporated in two new indices – the non-Gaussianity index (*NGI*) and the nonlinearity index (*NLI*). The magnitudes of new indices are always bounded between -1 to 1 because bicoherence, whose magnitudes are also bounded between 0 and 1, are used to define them. The definition and development of these indices using statistical test on a bounded function (bicoherence) allow the users to apply them for comparing multiple time series or control loops. The derivation and development of these indices are described below.

A discrete ergodic stationary time series, $x(k)$, is called linear if it can be represented by

$$x(k) = \sum_{n=0}^{M-1} h(n)e(k-n) \tag{3.1}$$

where $e(k)$ is a sequence of independent identically distributed random variables with

$E[e(k)] = 0$, $\sigma_e^2 = E[e^2(k)]$, and $\mu_3 = E[e^3(k)]$. For this case, the following frequency domain relationships can be obtained.

$$\text{The power spectrum:} \qquad P(f) = \sigma_e^2 \mid H(f) \mid^2 \equiv |X(f)X^*(f)| \tag{3.2}$$

$$\text{and the bispectrum:} \qquad B(f_1, f_2) = \mu_3 H(f_1)H(f_2)H^*(f_1 + f_2) \tag{3.3}$$

where $H(f) = \sum_{n=0}^{M-1} h(n)e^{-ifn}$. Equation 2.65 can be rewritten as

$$\begin{aligned} bic^2(f_1, f_2) &\triangleq \frac{|B(f_1, f_2)|^2}{E[|X(f_1)X^*(f_1)||X(f_2)X^*(f_2)|] \;\; E[|X(f_1 + f_2)X^*(f_1 + f_2)|]} \\ &\equiv \frac{|B(f_1, f_2)|^2}{E[|P(f_1)||P(f_2)|]E[|P(f_1 + f_2)|]} \end{aligned} \tag{3.4}$$

For a linear time series, substituting the expressions from Equation 3.2 and 3.3, it can be shown that

$$bic^2(f_1, f_2) = \frac{\mu_3^2}{\sigma_e^6} \tag{3.5}$$

Equation 3.5 shows that for any linear signal, x, the squared bicoherence will be independent of the bifrequencies, i.e., a constant in the bifrequency plane. If the squared bicoherence is zero, the signal x is Gaussian because the skewness or μ_3 is also zero in such a case. **Strictly speaking, such a signal should be called non-skewed with a symmetric distribution instead of a Gaussian one. However, in this thesis and also in most of the HOS literature (Nikias and Petropulu, 1993; Hinich, 1982; Rao and Gabr, 1980; Kim and Powers, 1979; Collis *et al.*, 1998; Fackrell, 1996) the two terms – nonskewed and Gaussian – have been used interchangeably**. To check whether the squared bicoherence is constant, two tests are required. One is to determine whether the squared bicoherence is zero, which would show that the signal is Gaussian and thereby the signal generating process is linear. The other is to test for a non-zero constant squared bicoherence which would show that the signal is non-Gaussian but the signal generating process is linear.

The bicoherence is a complex quantity with real and imaginary parts. The magnitude of the squared bicoherence can be obtained as

$$bic^2 = \Re(bic)^2 + \Im(bic)^2 \tag{3.6}$$

where $\Re$ and $\Im$ are real and imaginary parts, respectively. It is well well established in the HOS literature that bicoherence is a complex normal variable, i.e., both the estimates of real and imaginary parts of the bicoherence are normally distributed (Hinich, 1982) and asymptotically independent, i.e., the estimate at a particular bifrequency is independent

of the estimates of its neighboring bifrequencies (Fackrell, 1996). Therefore, the squared bicoherence at each frequency is a non-central chi-squared ($\chi^2(m)$) distributed variable with 2 degrees of freedom and mean, m. Note that bic^2 is bounded between 0 and 1 and therefore the $E\{bic^2\}$ and $var\{bic^2\}$ are also bounded between 0 and 1. Hinich (1982) showed that the signal of interest is Gaussian if the skewness function is asymptotically centrally χ^2 distributed with 2 degrees of freedom. This information was used by (Fackrell, 1996) to test bicoherence at each frequency in the principal domain. The disadvantage of this test is that while applying to each of the bifrequencies in the principal domain of squared bicoherence plot, the probability of false detection accumulates owing to a large number of bifrequencies in the principal domain. Thus it overestimates the number of bifrequencies in which the bicoherence magnitude is significant. A modified test with better statistical properties but no frequency resolution is formulated by averaging the squared bicoherence over the triangle of the principal domain. The test can be summarized as follows:

- Null Hypothesis, H_o: The signal is Gaussian
- Alternate Hypothesis, H_1: The signal is not Gaussian.

Under the null hypothesis, the test for the average squared bicoherence can be based on the following equation:

$$P(2KL\overline{\hat{bic}^2} > c_\alpha^{\chi^2}) = \alpha \tag{3.7}$$

where $c_\alpha^{\chi^2}$ is the critical value calculated from the central χ^2 distribution table for a significance level of α at 2L degrees of freedom since $\overline{\hat{bic}^2} = \sum_{i=1}^{L} \hat{bic}_i^2$ and L is the number of bifrequencies inside the principal domain of the bispectrum, K is the number of data segments used in the bicoherence estimation.

If the number of bifrequencies in the principal domain is large (more than 100) the normal approximation of the χ^2 distribution can be used. The approximation is given by Abramowitz and Stegun (1972):

$$c_\alpha^{\chi^2} = \frac{1}{2}[c_\alpha^z + \sqrt{2dof - 1}]^2 \tag{3.8}$$

where $c_\alpha^{\chi^2}$ and c_α^z are the critical values of χ^2 and standard normal distribution at a significance level of α, respectively and dof is the degrees of freedom. Now, substituting Equation 3.8 into 3.7 with 2L degrees of freedom, it can be shown that

$$P(\overline{\hat{bic}^2} > \frac{1}{4KL}[c_\alpha^z + \sqrt{4L - 1}]^2) = \alpha \tag{3.9}$$

This equation can be rewritten as

$$P(\overline{\hat{bic}^2} - \overline{bic^2_{crit}} > 0) = \alpha \tag{3.10}$$

$$or, \quad P(NGI > 0) = \alpha \tag{3.11}$$

where $\overline{bic^2_{crit}} = \frac{1}{4KL}[c^2_\alpha + \sqrt{4L-1}]^2$ and $NGI \triangleq \overline{\hat{bic}^2} - \overline{bic^2_{crit}}$ NGI stands for the Non-Gaussianity Index and it is bounded between -1 to 1. Therefore, at a confidence level of α, the following rule-based decision can be obtained:

- **if $NGI \leq 0$, the signal is GAUSSIAN**
- **if $NGI > 0$, the signal is NON-GAUSSIAN**

Thus, a signal is Gaussian (non-skewed) at a confidence level of α if the NGI is less than or equal to zero. This index has been defined to facilitate the automation of the decision.

If the signal is found to be Gaussian, the signal generating process is assumed to be linear. In the case of non-Gaussian signals the signal generating process should be tested for its linearity. As shown in Equation 3.5, if the signal is non-Gaussian and linear, the magnitude of the squared bicoherence should be a non-zero constant at all bifrequencies in the principal domain. The constancy of the squared bicoherence (skewness) was tested using an F test by Rao and Gabr (1980). Hinich (1982) reported that this test is highly vulnerable to outliers. He suggested a method based on the Sample Interquartile Range (SIQR) of the χ^2 distribution of the squared skewness function (Hinich, 1982). However, this test depends on the sample size of the time series. Also, the SIQR is not an ideal measure for constant bicoherence because it is readily seen that the SIQR of the squared bicoherence can be zero, though all squared bicoherence values are not equal (Yuan, 1999). A simple way to confirm the constancy of squared bicoherence is to have a look at the 3-D squared bicoherence plot and observe the flatness of the plot. However, this can be tedious and cumbersome for a large number of loops. Alternatively, if the squared bicoherence is of a constant magnitude at all bifrequencies in the principal domain, the variance of the estimated bicoherence should be zero. To check the flatness of the plot or the constancy of the squared bicoherence, the maximum squared bicoherence can be compared with the average squared bicoherence plus two or three times the standard deviation of the estimated squared bicoherence. At a 95% confidence level, if the maximum squared bicoherence, $\hat{bic}^2_{max}$, is less than $(\overline{\hat{bic}^2} + 2\sigma_{\hat{bic}^2})$, then the magnitudes of squared bicoherence are assumed to be a constant or the surface is flat. The automatic detection of this can be performed using the following Nonlinearity Index (NLI), which is defined as:

$$NLI \triangleq \hat{bic}^2_{max} - (\overline{\hat{bic}^2} + 2\sigma_{\hat{bic}^2}) \tag{3.12}$$

where $\sigma_{\hat{bic}^2}$ is the standard deviation of the estimated squared bicoherence and $\overline{\hat{bic}^2}$ is the average of the estimated squared bicoherence. Ideally, the *NLI* should be 0 for a linear process. This is so because, if the squared bicoherence is a constant at all frequencies, then the variance will be zero and both the maximum and the mean will be same. Therefore, it can be concluded that:

- **if $NLI \leq 0$, the signal generating process is LINEAR**
- **if $NLI > 0$, the signal generating process is NONLINEAR**

Since the squared bicoherence is bounded between 0 and 1, the Nonlinearity Index (*NLI*) is also bounded between -1 and 1.

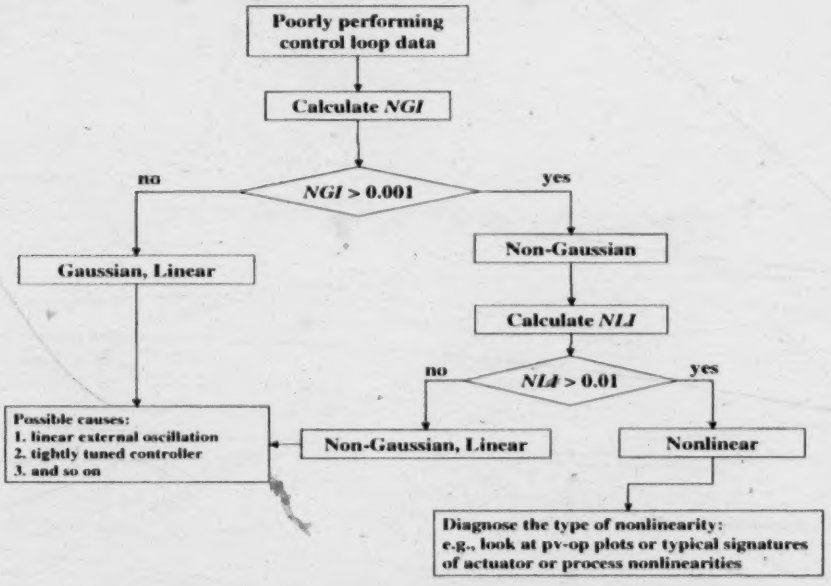


Figure 3.3: *Rule based decision flow diagram.*

Practical Implementation: In practice, it is difficult to obtain an exact zero value of *NGI* for Gaussian signals. Therefore, we select a threshold value, ε, of *NGI* such that $NGI < \varepsilon$ implies a Gaussian signal. To the best knowledge and experience of the author, for $\alpha = 0.05$, an *NGI* value of less than $\varepsilon = 0.001$ can be assumed to be zero. Consequently, if $NGI \leq 0.001$, the signal can be assumed to be Gaussian at a 95% confidence level. For *NLI*, a value less than 0.01 is assumed to be zero and consequently, the process is considered to be linear at a 95% confidence level. The larger the value of *NLI*, the higher the extent of

nonlinearity. The detailed diagnosis procedure can be summarized in a rule-based decision flow diagram shown in Figure 3.3.

3.4 Illustrative Examples

3.4.1 Bicoherence of a linear and nonlinear signal

Two signals, y_{linear} and $y_{nonlinear}$, were generated using the following equations.

$$\begin{aligned} x(k) &= 3H(q^{-1})d_1(k) \\ y_{linear} &= x(k) + d_2(k) \qquad (3.13) \\ y_{nonlinear} &= x(k) + 0.1x(k)^2 + d_2(k) \qquad (3.14) \end{aligned}$$

where $d_1(k)$ and $d_2(k)$ are zero mean white noises with variance 1 and 0.001 respectively, and $H(q^{-1})$ is a narrow pass Butterworth filter with a frequency range 0.095 to 0.105 in a 0 to 0.5 normalized frequency scale, such that $f_s = 1$ is the sampling frequency.

The purpose of this example is to demonstrate the power of the bicoherence in the detection of nonlinearity. By merely looking at the time trend of the signals (the left panel of Figure 3.4), it is not possible to differentiate between them. Also, the power spectrums (the middle panel of Figure 3.4) or the second order moments look alike and are unable to detect the nonlinearity present in the second signal. The right panel of Figure 3.4 shows the three dimensional bicoherence plots. For y_{linear}, the test result is $NGI = -0.0028$. Clearly, the NGI indicates that the signal is Gaussian. Therefore, the nonlinearity test result is not required here. In contrast, for $y_{nonlinear}$, the NGI equals 0.002, thereby detecting the non-Gaussianity of the signal. The nonlinearity test gives $NLI = 0.37$, which clearly indicates the presence of nonlinearity in this signal. From the bicoherence plot, the peak position in the principal domain is approximately at the (0.1,0.1) bifrequency. This means that the nonlinearity in the signal is due to the interaction of these two frequencies. Examination of the signal generating system reveals that the band-pass filtered signal has the frequency range [0.095 to 0.105]. This signal was squared to introduce nonlinearity. Thus, the nonlinearity is due to the multiplication of two signals, each having a frequency of approximately 0.1. In the bicoherence plot the bifrequency of the identified peak is (0.1,0.1). Therefore, the HOS based method correctly identifies the frequencies of nonlinear interactions.

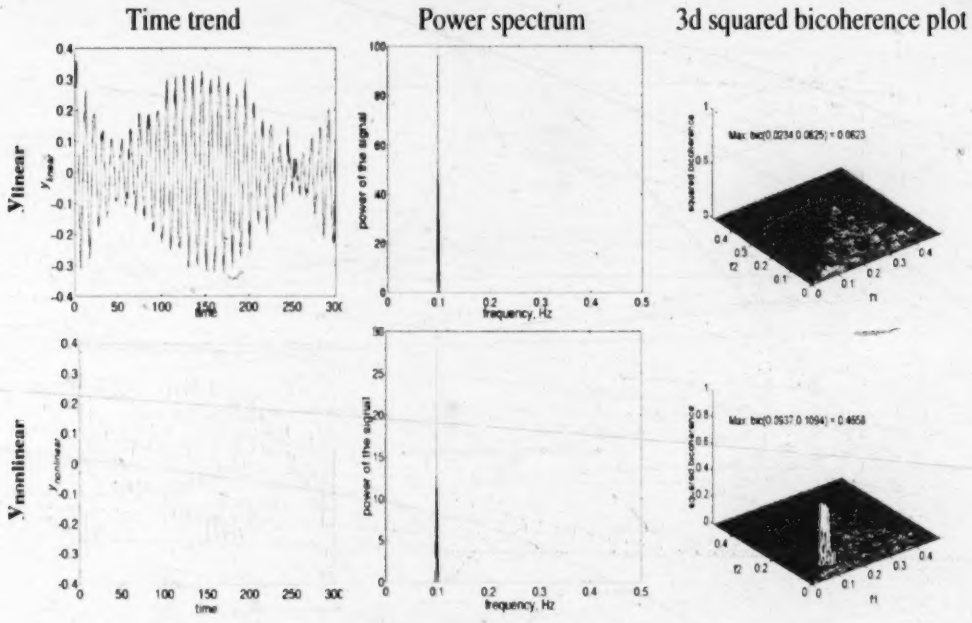


Figure 3.4: *Results for y_{linear} (top) and $y_{nonlinear}$ (bottom).*

3.4.2 Bicoherence of a nonlinear sinusoid signal with noise

An input signal was constructed by adding two sinusoids, each having a different frequency and phase. That is,

$$\begin{aligned} x'(k) &= sin(2\pi f_1 k + \phi_1) + sin(2\pi f_2 k + \phi_2) \\ x(k) &= x'(k) + d(k) \\ y(k) &= x'(k) + n_l x'(k)^2 + d(k) \end{aligned} \tag{3.15}$$

where, $f_1 = 0.12$, $f_2 = 0.30$, $\phi_1 = \pi/3$, $\phi_2 = \pi/8$, n_l is a multiplication factor employed to represent the contribution of the nonlinear component of the signal, and $d(k)$ is a white noise sequence with variance 0.04. Again, frequencies are normalized such that the sampling frequency is 1.

The quadratic term in Equation 3.15 will introduce phase coupling in the output signal,

Table 3.1: QPC relations for the output signal, y

	frequency relations	phase relations	peak locations
1	$f_1+f_1=2f_1$	$\phi_1+\phi_1=2\phi_1$	(f_1,f_1)
2	$f_2+f_2=2f_2$	$\phi_2+\phi_2=2\phi_2$	(f_2,f_2)
3	$f_1+f_2=f_1+f_2$	$\phi_1+\phi_2=\phi_1+\phi_2$	(f_1,f_2)
4	$f_1+(f_2-f_1)=f_2$	$\phi_1+(\phi_2-\phi_1)=\phi_2$	(f_1,f_2-f_1)
5	$(f_2-f_1)+(f_1+f_2)=2f_2$	$\phi_1+\phi_2+(\phi_2-\phi_1)=\phi_1+\phi_2$	(f_2-f_1,f_1+f_2)
6	$(f_2-f_1)+2f_1=f_1+f_2$	$2\phi_1+(\phi_2-\phi_1)=\phi_1+\phi_2$	$(f_2-f_1,2f_1)$

y. It can be better understood by rewriting Equation 3.15 in the following form:

$$\begin{aligned} y(k) &= \sin(2\pi f_1 k+\phi_1)+\sin(2\pi f_2 k+\phi_2)+0.25[1-\cos(2(2\pi f_1 k+\phi_1)) \\ &\quad -\cos(2(2\pi f_2 k+\phi_2))+\cos(2\pi(f_2-f_1)k+\phi_2-\phi_1) \\ &\quad -\cos(2\pi(f_1+f_2)k+\phi_1+\phi_2)]+d(k) \end{aligned} \tag{3.16}$$

The nonlinearities can be generated by the interactions of any two of the signals with frequencies f_1, f_2, $2f_1$, $2f_2$, f_2-f_1, and f_1+f_2. Due to the presence of the quadratic nonlinearity function, there can be six frequency couplings. These are presented in Table 3.1.

1. **Case 1: Mild Nonlinearity** ($n_l = 0.05$)

   The left panel of Figure 3.5 shows the time series while the middle panel depicts the power spectrum of the signal x and y, respectively. Neither of these plots help in distinguishing the two signals. However, the use of higher order statistics can successfully detect the nonlinearities present in y. The right panel of Figure 3.5 shows the three-dimensional squared bicoherence plots of x and y, respectively. For the signal x, $NGI = 0.0008$, which clearly indicates that the signal is Gaussian and linear.

   In contrast, in the case ofr y the test results are $NGI = 0.016$ and $NLI = 0.233$. Thus the nonlinearity present in y is correctly detected. This example also shows the sensitivity of the proposed indices to the presence of nonlinearity in the signal. The presence of as little as 5% of the nonlinear square term in the signal y has been detected. For this case, the squared bicoherence plot shows

peaks at (0.12,0.12), (0.12,0.18), (0.30,0.30), and (0.12,0.30) bifrequencies. These bifrequencies correspond to (f_1, f_1),$(f_1, f_2 - f_1)$, (f_2, f_2) and (f_1, f_2) respectively. Note that because only 5% of the nonlinear term was added, the other two peaks are not visible in the bicoherence plot owing to their small sizes.

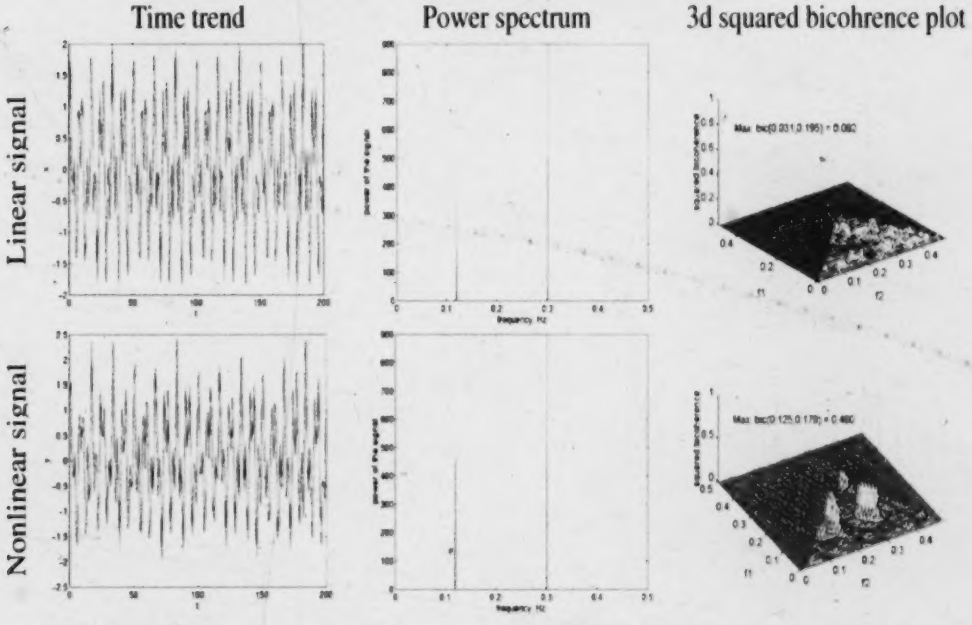


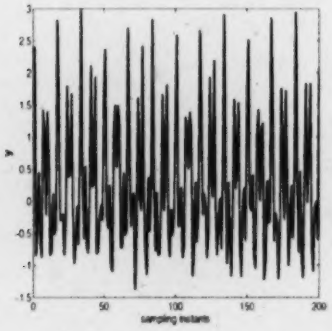
Figure 3.5: *HOS analysis results for the linear and nonlinear sinusoid signals, case 1 - mild nonlinearity.*

2. **Case 2: Strong nonlinearity** ($n_l = 0.25$)

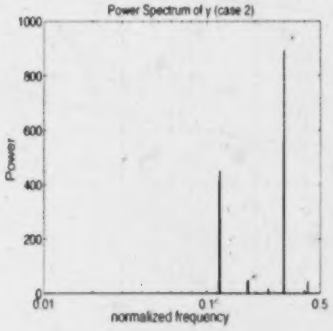
   For this case, the signal y was produced using the same Equation 3.15 but with a larger multiplying factor for the nonlinear quadratic term in order to observe all the peaks in the bicoherence plot resulting from the theoretical analysis in Table 3.1. The magnitude of n_l was chosen as 0.25. A different value of n_l can also be picked.

   Figure 3.6(a) shows the time trend of y, while the power spectrum of y is shown in Figure 3.6(b). This time power spectrum shows extra peaks of small magnitude at frequencies 0.18. 0.24, and 0.42. The use of higher order statistics can successfully detect the nonlinearities present in y. The magnitudes of *NGI* and *NLI* are 0.02

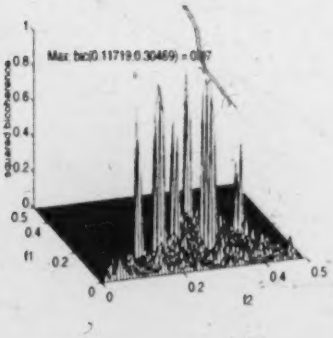
and 0.67, respectively, clearly detecting the nonlinearity of the signal. Figure 3.6(c) depicts the three-dimensional squared bicoherence plots for y. In order to visualize the location of the bicoherence peaks, a contour plot for the bicoherence magnitude is presented in Figure 3.6(d). This plot clearly shows the location of all six peaks at (0.12,0.12), (0.30,0.30), (0.12,0.30), (0.12,0.18), (0.42,0.18), and (0.18,0.24) bifrequencies. These bifrequencies correspond to (f_1, f_1), (f_2, f_2), (f_1, f_2), $(f_1, f_2 - f_1)$, $(f_1 + f_2, f_2 - f_1)$, and $(f_2 - f_1, 2f_1)$ respectively. Note that since 25% of the nonlinear term was added, all the peaks were visible for this case. Therefore, the bicoherence plot correctly identifies the frequency interactions resulting from the presence of nonlinearity in the signal.



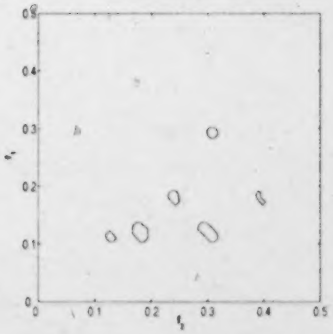
(a) time trend of y



(b) power spectrum of y



(c) bicoherence plot



(d) bicoherence contour plot

Figure 3.6: *HOS analysis results for case 2 - strong nonlinearity.*

3.5 Investigation of Control Loop Nonlinearities

In a control loop, nonlinearities may appear in several locations:

1. The process itself may be nonlinear in nature.
2. The control valve may have a nonlinear characteristic.
3. The valve may be affected by some nonlinear faults, e.g., stiction, deadband and hysteresis.
4. A nonlinear disturbance may enter the loop.

In order to detect and diagnose any problem related to loop nonlinearities all of the above should be investigated carefully. In this section, each of these nonlinear sources is examined separately.

3.5.1 Process nonlinearity

It is often assumed that the process operates in a locally linear fashion. This section investigates this assumption using results obtained from the application of higher order statistical tests. Analysis of process nonlinearity using HOS will be illustrated using an example obtained from Agrawal and Lakshminarayanan (2003). Consider the control of water level in a spherical tank by manipulating the inlet volumetric flow rate, F_i (see Figure 3.7). The dynamics of the system can be modelled as:

$$\pi R^2[1-\frac{(R-h)^2}{R^2}]\frac{dh}{dt} = F_i(t-d) - F_o(t) \tag{3.17}$$

where $F_o(t)$ is the outlet flow rate at time t, h is the height of the water level from the bottom of the tank, R is the radius of the spherical tank, and d is the time delay between F_i and F_o. The outlet flow rate, F_o, can be expressed as

$$F_o(t) = \sqrt{2g(h-h_o)} \tag{3.18}$$

where g is the gravitational constant and h_o is the height of the outlet pipe from the base of the vessel. For both open loop and closed loop simulations of the system, we use $R = 0.5$ m, $h_o = 0.01$ m, and the nominal operating point is selected to be, $h_{ss} = 0.30$ m.

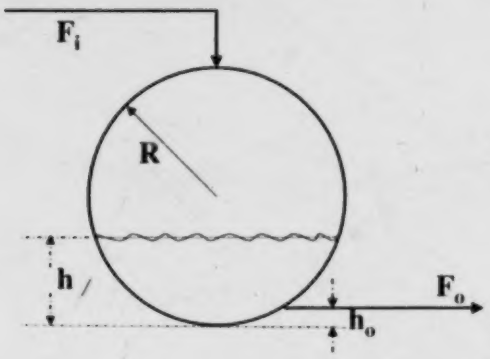
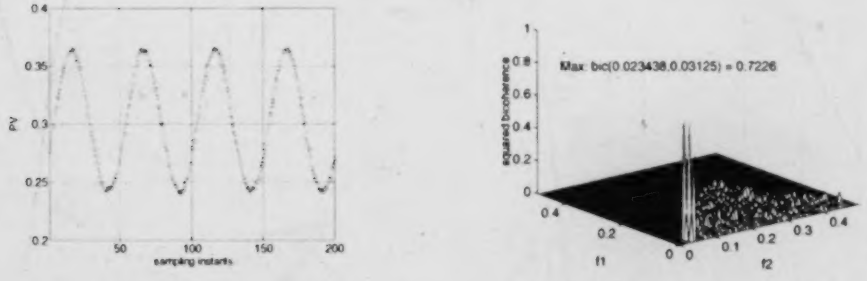


Figure 3.7: *Level control of a spherical tank.*



(a) time trend of process output

(b) 3d squared bicoherence

Figure 3.8: *Results of the analysis of open loop data for the spherical tank system.*

Open Loop Simulation of the Process

The input to the process was a sinusoid with frequency 0.02 Hz and amplitude 0.25. 4096 data points were used for nonlinear analysis. The nonlinearity test results in $NGI = 0.0024$ and $NLI = 0.633$. This clearly demonstrates that the process is nonlinear. The time trend of *pv* and the squared bicoherence plot are shown in Figure 3.8. Since the excitation was very large ($2 * 0.25/1 = 50\%$ *of the tank height*) the nonlinearity of the process could not be ignored. The nonlinearity of any process strongly depends on the size of the excitation or input signal. For example, if the same spherical tank process is excited by a sinusoid with an amplitude of 0.05, the process does not show any nonlinearity in the bicoherence test. The values of *NGI* and *NLI* are -0.001 and 0.05. The corresponding squared bicoherence

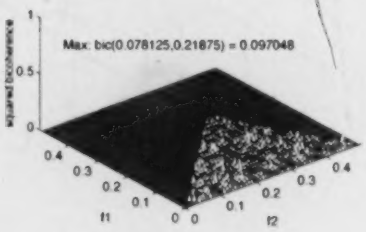


Figure 3.9: *Squared bicoherence for small excitation of spherical tank level.*

plot is shown in Figure 3.9. Therefore, a system with small excitation can be assumed to be locally linear.

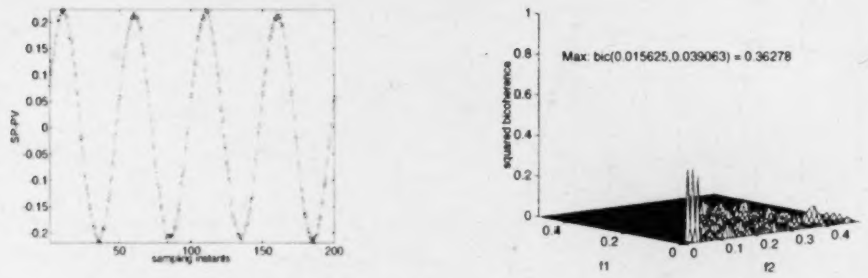
Closed Loop Simulation of the Process

The PI controller settings used for this simulation are given by:

$$C(z^{-1}) = \frac{0.3 - 0.1z^{-1}}{1 - z^{-1}} \tag{3.19}$$

The input to the process was a sinusoid with frequency 0.02 Hz and amplitude 0.25. Four thousands and Ninety Six (4096) data points from the controller error signal were used for nonlinear analysis. The nonlinearity test results in $NGI = 0.0011$ and $NLI = 0.30$. This clearly detects that the process is nonlinear. The time trend of $sp - pv$ and the squared bicoherence plot are shown in Figure 3.10. Large peaks are observed in the bicoherence plot. If the magnitude of the input excitation signal or set point is changed to 0.05, the nonlinearity test applied to the error signal gives $NGI = -0.01$, indicating a linear process. The flatness of the bicoherence plot shown in Figure 3.11 shows that the process can indeed be considered locally linear.

Remark: *The assumption of linearity of a process is case-dependent. Most control loops in chemical process industries operating under regulatory feedback control for rejecting disturbances usually have small excitations. Therefore, they can be assumed to be locally linear.*



(a) time trend of controller error

(b) squared bicoherence plot

Figure 3.10: *Results of the analysis of closed loop data for the spherical tank system.*

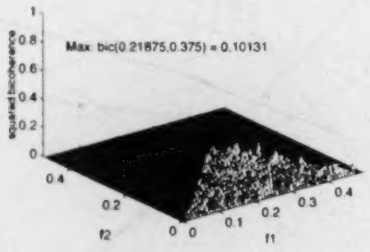


Figure 3.11: *Squared bicoherence for small excitation of the spherical tank level under feedback control. The flatness of the plot shows the linearity of the process.*

3.5.2 Nonlinear valve characteristic

The valve is the final control element in a control loop. The flow through a valve is often described by the following relationship:

$$Q = KC_v(x)\sqrt{\frac{\Delta P_v}{s.g.}} \tag{3.20}$$

where Q is the volumetric flow rate through the valve, K is the constant that depends on the units used in the equation, C_v is the valve co-efficient which depends on the inherent valve characteristics and the stem position (x), ΔP_v is the pressure drop across the valve, and $s.g.$ is the specific gravity of the fluid. There are three common valve characteristics:

- Linear characteristics: $C_v(x) = x$
- Equal percentage valve: $C_v(x) = \kappa^{(x-1)}$
- Quick opening or square root valve: $C_v(x) = \sqrt{x}$

where, κ is constant and 50 is a value frequently used for it.

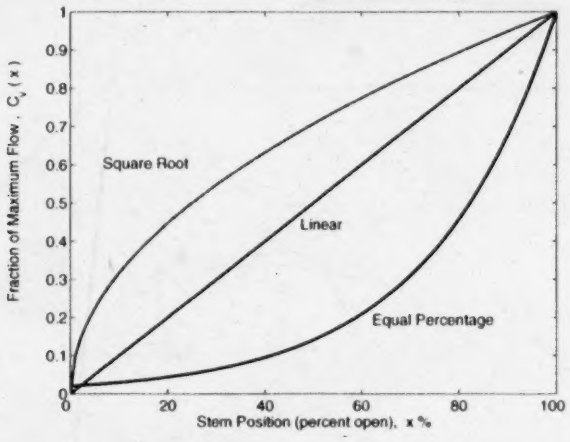


Figure 3.12: *Inherent characteristics for linear, equal percentage and square-root valve.*

These valve characteristics are illustrated in Figure 3.12. Note that for quick-opening valves, the sensitivity of the flow to the fraction opening or stem position is high at a low flow rate and low at a high flow rate. The opposite is true for equal percentage valves. These relationships have been defined for constant pressure drop across the valve carrying an incompressible fluid. Thus, these relationships are valid only under such conditions.

Linear valves

As the name suggests, a linear valve has linear characteristics. There is no inherent nonlinearity in this type of valve. The other two types of valves will be examined for their nonlinear characteristics in the following section.

Equal percentage valves

These are called equal percentage valves because an equal incremental increase in valve travel or stem position causes a constant percentage increase of flow from the earlier flow.

For example, if a value of $\kappa = 50$ is used, for every 10% increase in stem position or valve travel x, there will be an increase of 47.88% of the flow from the previous flow. The effect of nonlinearity in the characteristic equation of this type of valve has been investigated using the following simulation example.

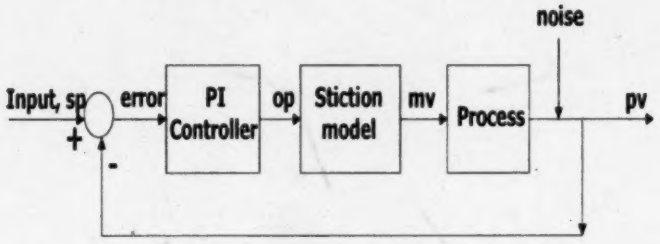


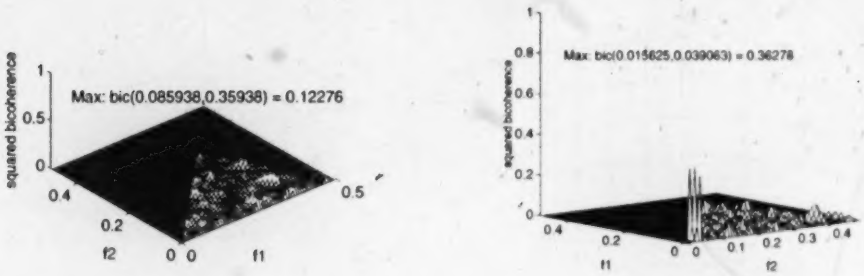
Figure 3.13: *Block diagram of a simple SISO process with stiction nonlinearity in the valve.*

A simple single-input, single-output (SISO) system in a feedback control configuration (Figure 3.13) was used to generate simulated data. The first order process with time delay is given by the following transfer function:

$$G(z^{-1}) = \frac{z^{-3}(1.45 - z^{-1})}{1 - 0.8z^{-1}} \tag{3.21}$$

The process is regulated by a PI controller. A random white noise with a variance of 0.5 was added to the process. The simulation was performed for a steady state of 12 mA or the valve position at 50%. The valve had to travel in the range of 25% around the steady state in order to reject the disturbances to the loop. The PI controller parameters for this case were $K_c = 0.15$ and $I = K_c/\tau_i = 0.15\ second^{-1}$. The nonlinear 'stiction model' block was removed from the simulation block diagram and a function describing the equal percentage valve characteristic was placed in that position. The simulation was performed for 6000 sampling intervals. To remove the effect of transients, the first several hundred data points were discarded, and then the last 4096 points of the error signal to the controller (*sp-pv*) were analyzed to detect the nonlinearity present in the system. Transients have a serious contamination effect on the estimated bicoherence. For details, refer to (Fackrell, 1996). The proposed nonlinearity test shows that $NGI = -0.0011$. This shows that the error signal is Gaussian and linear indicating linearity of the loop. The left plot of Figure 3.14 shows the squared bicoherence plot for the control error of this loop.

When the noise variance was increased to 1.5, the valve had to travel 40% of its full range. The nonlinearity test shows that NGI=0.003 and NLI=0.07. This indicates the loop



(a) squared bicoherence plot

(b) squared bicoherence plot

Figure 3.14: *Bicoherence plot for an equal percentage valve. Left - Valve travels less than 25% of its full span. Right- Valve travels more than 25% of its full span.*

is nonlinear. The corresponding bicoherence plot is shown in the right plot of Figure 3.14.

Square-root valve

The quick opening or square root valve provides large changes in flow for small changes in valve travel or stem position. Therefore, it has a high valve gain unsuitable for use in regulatory control. Its use is limited to on-off service. To investigate the nonlinear behavior of this valve, a simulation study was performed using the procedure described in the case of the equal percentage valve, but with the use of a square root valve instead of an equal percentage one. The valve was forced to travel in the range of 17.8% around the steady state (50%) to eliminate the effect of disturbances. As before to remove the effect of transients, the first several hundred data points were discarded. The last 4096 points of the error signal to the controller (*sp-pv*) were analyzed to detect the nonlinearity present in the system. The proposed nonlinearity test shows that $NGI = -0.0010$. This indicates that the error signal is Gaussian and linear indicating the linearity of the loop.

Remarks on nonlinear valve characteristic

As demonstrated, if the valve travels within a small range of the whole travel span (0-100%), say within 20%, then the valve characteristics can be assumed linear and does not add any nonlinearity to the loop. This observation can also be realized by investigating the valve characteristic curves in Figure 3.12. For any $\Delta x = 20\%$, the characteristic curves can

be assumed to be linear. These characteristic curves are called 'inherent characteristics' of the valve. They are only valid for constant pressure drop across the valve and for an incompressible fluid flow. However, in real life the pressure drop across the valve does not remain constant. Moreover, the valve is connected to other process equipment. Therefore, the characteristic curves of the valve should be re-evaluated after installing the valve in a real process. These characteristics are termed 'installed characteristics'.

In reality, the flow through the valve also depends on the pressure drop across the valve. Manufacturers test valves in a rig, where the pressure drop is kept constant. Thus the performance they observe is the 'inherent characteristics' of the valve. In a real plant, pressure drop varies as the flow changes. Therefore, the characteristic relationship seen between valve travel and flow will not be the same as that seen in the test rig. This 'installed characteristic' is what really matters to a process engineer. That is why during the selection and design of a valve, considerable effort is made to ensure that the 'installed characteristics' of the valve is as linear as possible (Fisher-Rosemount, 1999; Fitzgerald, 1995; Baumann, 1994; Riggs, 1999). As described in (Fisher-Rosemount, 1999), it is a good operating practice to keep valve swings below 5% so that the loop gain does not change much and stability of the loop is ensured. Based on the author's experience in industrial data analysis, it has been found that the controller output in most cases swings below 10%. Therefore, the control valve characteristics can be assumed to be locally linear in most cases of real world data analysis.

3.5.3 Nonlinear faults in valve

Control valves are mechanical devices subject to wear and tear with time. That is why they require regular maintenance. With time, they may develop various problems such as a large deadband, excessive static friction or stiction, saturation, backlash and seat corrosion. All these problems reduce performance of the control loop. Detection and diagnosis of such problems are the main objective of this thesis. The problem of valve stiction has been studied in detail herein.

Valve Saturation

All equipment has physical limits. A valve can not open more than 100% or can not close more than 0%. If the controller output demands a valve be open more than 100% or close more than 0%, the valve saturates. The input-output behaviour of saturation is shown in Figure 3.15. If there is persistent saturation in a control valve, control performance

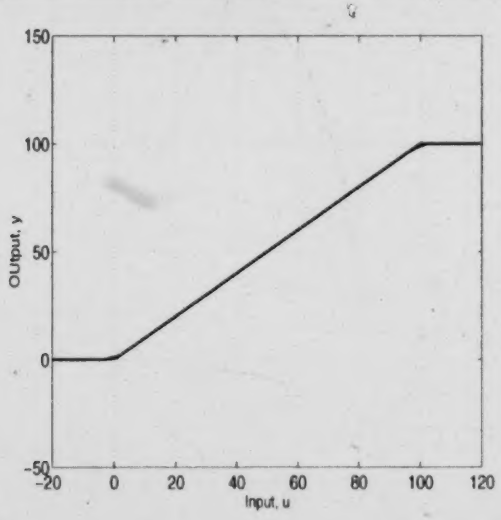


Figure 3.15: *Input-output characteristics of valve saturation.*

deteriorates and oscillations originate in the process variable. Therefore, techniques for diagnosing poor performance must be capable of detecting valve saturation.

3.5.4 Nonlinear disturbances

The disturbance entering a control loop may be linear or nonlinear. If the disturbance is measurable, the nonlinearity test developed in Section 3.3 can be applied to assess its nonlinearity. If the disturbance is unmeasurable, it is difficult to infer anything about the linearity of the loop. For the sake of simplification, the unmeasured disturbances entering the control loops are assumed to be linear in this study. More discussions on it are provided in chapter 5.

3.6 Simulation Example to Diagnose the Causes of Poor Performance

Poor performance of a control loop can be caused by various factors. Examples include (1) poorly tuned controllers, (2) the presence of oscillatory disturbances and (3) nonlinearities. The purpose of this simulation is to demonstrate the application of HOS based techniques

in diagnosing causes of poor performance. If the method does not detect nonlinearity, then the focus of the diagnosis should be on controller tuning or on the possible presence of an external linear oscillatory disturbance. If the method detects nonlinearity, then the nonlinearity must be isolated or localized. Is it in the valve or in the process? This study assumes that the process is locally linear.

The system described in the Section 3.5.2 has also been used for this simulation study. The process is regulated by a PI controller. An integrated white noise generated by integrating random noise with a variance of 0.05 was added to the process. The simulation was performed for 6000 sampling intervals. To remove the effect of transients, the first several hundred data points were discarded. The last 4096 points of the error signal to the controller (*sp-pv*) were analyzed to detect the nonlinearity present in the system.

3.6.1 Well tuned controller

The PI controller parameters for this case were $K_c = 0.15$ and $I = K_c/\tau_i = 0.15/1 = 0.15$ $second^{-1}$. The nonlinear 'stiction model' block was removed from the simulation block diagram. The top row of Figure 3.16 shows the results for this case. The proposed test shows that $NGI = -0.0008$. This indicates that the error signal is Gaussian and linear. The corresponding bicoherence plot is flat.

3.6.2 Tightly tuned controller

For this case, the controller parameters were set at $K_c = 0.15$ and $I = K_c/\tau_i = (0.15/0.4) = 0.375$ $second^{-1}$. The second row of Figure 3.16 shows the results. The presence of relatively large integral action produces large oscillations in the process variables. An *NGI* value of -0.0007 indicates the Gaussianity and linearity of the system. It suggests that the poor performance is not due to nonlinearities. Also, since there are no external oscillatory disturbances, the probable cause is the presence of a tightly tuned controller.

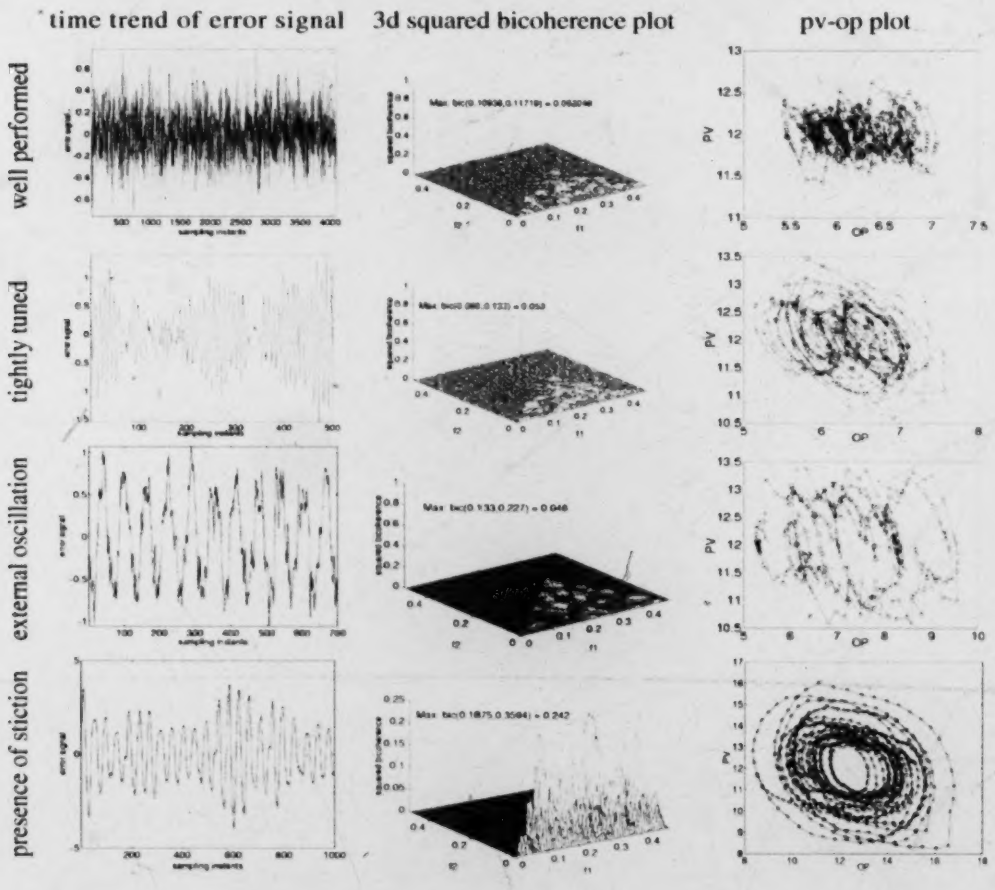


Figure 3.16: *Results of the simulation example. Bicoherence correctly detects the first three cases as linear and the last case as nonlinear. The pv-op plot for the stiction case shows elliptic patterns.*

3.6.3 Presence of an external oscillatory disturbance

A sinusoid with amplitude 2 and frequency 0.01 was added to the process output in Figure 3.13 in order to feed an external oscillatory disturbance to the process. The results for this case are shown in the third row of Figure 3.16. Horch's correlation method (Horch, 1999) of diagnosing the oscillation or more specifically valve stiction gives an odd correlation

function between *op* and *pv* for this case, thereby falsely detecting the presence of stiction in the control loop (more discussion will be provided in Section 5.6.2). The proposed test gives an *NGI* value of -0.0003. It clearly shows that the reason for the oscillation is not due to presence of nonlinearity in the system. The bicoherence plot is also flat.

3.6.4 Presence of stiction

A valve stiction model developed as part of this work (see Chapter-4, plus Choudhury and Shah (2001) and Choudhury *et al.* (2004*a*)) was used to perform this simulation. The model consists of two parameters: (1) deadband plus stickband, S and (2) slip jump, J. Figure 4.11 in Chapter 4 summarizes the model algorithm.

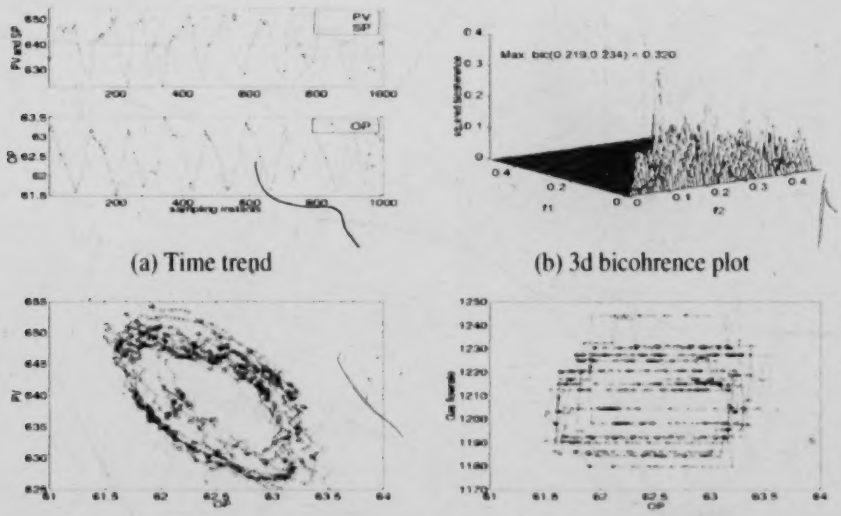
To perform the simulation for this particular case, $S = 3$ and $J = 1$ were used. Note that in order to initiate limit cycles or oscillations in a simple first order time delay process in the presence of valve stiction, a set point change at the beginning of the simulation is required. Thereafter, the process is allowed to operate under regulatory control. The last row of Figure 3.16 shows the time trend of the control error signal, the bicoherence plot and the *pv-op* plot. The presence of stiction produces oscillations in the process. The values of *NGI* and *NLI* are 0.05 and 0.048 (the thresholds for *NGI* and *NLI* are 0.001 and 0.01 respectively), clearly detecting the presence of nonlinearity in the process signal. After detecting the nonlinearity, the process variable versus controller output plot, i.e., *pv-op* plot can be used to diagnose the type of nonlinearity. Usually, the presence of distinct elliptical loops with sharp-turn around points is an indication of the presence of stiction in the valve. Note that for other cases there are no such distinct cycles in *pv-op* plot (see the right panel of the Figure 3.16). If the valve position (*mv*) is available, the *mv-op* plot can be conveniently and more accurately used to identify the type of nonlinearity in the valve.

3.7 Industrial Case Studies

The proposed method has been successfully applied to the detection and isolation of actuator or valve faults for many industrial control loops. There are many types of valve faults such as stiction, saturation, oversized valve, and corroded valve seat. One diagnosis example for each type of valve faults will be discussed here.

3.7.1 Stiction in a furnace dryer temperature control valve

This subsection describes a temperature control loop on a furnace feed dryer system at the Tech-Cominco mine in Trail, British Columbia, Canada. The temperature of the dryer combustion chamber is controlled by manipulating the flow rate of natural gas to the combustion chamber. The minimum variance performance index of this loop was very poor.



(a) Time trend

(b) 3d bicohrence plot

(c) pv-op characteristic plot

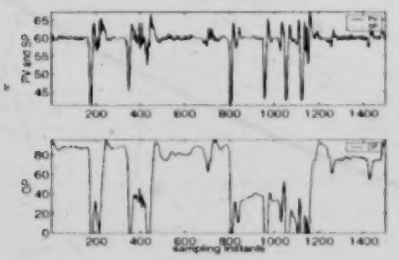
(d) mv (flow rate) – op plot

Figure 3.17: *Analysis of time series data from an industrial temperature control loop.*

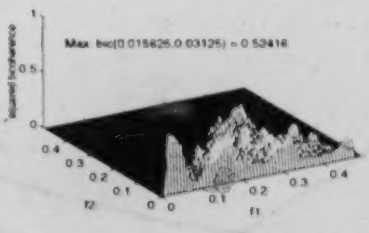
Figure 3.17(a) depicts time trends of the controlled variable, the controller output and the set point. It shows clear oscillations in both the controlled variable and the controller output. Figure 3.17(b) shows the bicoherence plot. The *NGI* and *NLI* obtained for this loop were 0.006 and 0.197 respectively, clearly indicating the presence of nonlinearity in the loop. The presence of distinct cycles in the characteristic *pv-op* plot (Figure 3.17(c)) together with the pattern obtained in Figure 3.17(d) characterize the presence of backlash and stiction in the control valve. Thus, this analysis was able to confirm the cause of poor loop performance due to the presence of valve stiction.

3.7.2 Valve Saturation

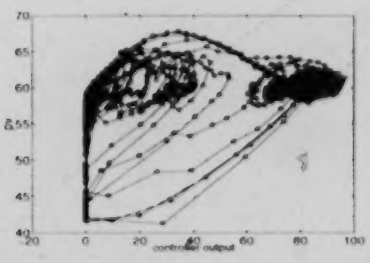
Figure 3.18 illustrates an example of an industrial control loop where the valve suffers from saturation. The time trends of the *pv*, *op*, and *sp* are depicted in Figure 3.18(a) showing oscillations. The bicoherence test gives $NGI = 0.07$ and $NLI = 0.29$, clearly indicating a nonlinear loop. The corresponding bicoherence plot also shows many large peaks indicating significant nonlinearity. To diagnose the type of nonlinearity, specific patterns in *pv-op* plot are found to be useful. The *pv-op* mappings for this loop is shown in Figure 3.18(c). A vertical straight line with some random cycles in the *pv-op* plot is a signature of valve saturation. If the valve saturates in both ends (that is for both full close and full open conditions), there will be two vertical straight lines in the *pv-op* plot. The explanation of this pattern lies in the use of anti-wind up in the integral action of the PI(D) controller. Because of anti-wind up action, the controller output is kept constant while the process output may change.



(a) time trends



(b) squared bicoherence plot

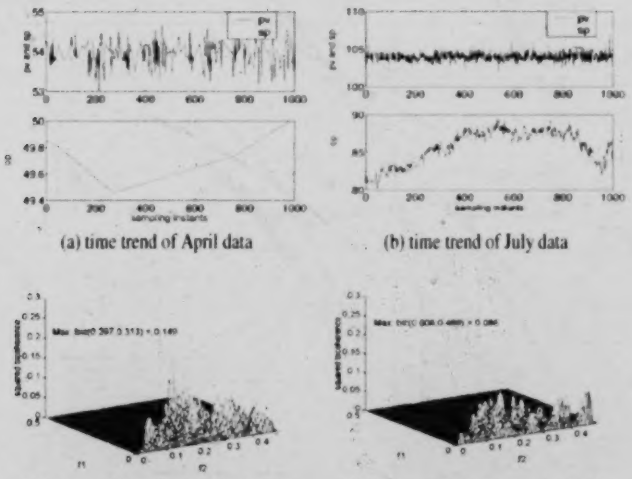


(c) pv-op plot

Figure 3.18: *Results of the analysis of an industrial loop with valve saturation.*

3.7.3 Valve problems in some flow control loops

This analysis is for two flow control loops at Celanese Canada Ltd., a chemical complex located in Edmonton, Canada. Data were collected with a sampling interval of 1 min over two periods of time: April 10 to 17, 2001 and July 1 to 15, 2001, the latter following an annual maintenance shutdown of the plant. Results for both these loops are discussed below.



(a) time trend of April data (b) time trend of July data

(c) 3d squared bicohrence plot (April) (d) 3d squared bicohrence plot (July)

Figure 3.19: *Analysis of flow loop 1 data before (April - left) and after the (July - right) plant maintenance shutdown period.*

Oversized valve

This is a recycle flow control loop. Detailed diagnostic plots are shown in Figure 3.19. Time series of the data collected in April and July are shown in Figures 3.19(a) and 3.19(b). The *op* trend in Figure 3.19(a) shows that the valve movement was very slow and insignificant compared to the change in the error signal, (*pv-sp*). The values of *NGI* and *NLI* are 0.01 and 0.10 respectively. An *NGI* value of 0.01 shows that the signal is Non-Gaussian. The *NLI* value of 0.10 indicates the presence of nonlinearity in the error signal. The *op* time trend in Figure 3.19(a) shows that a small change in *op* caused a big change in *pv* value (note the range of y-axis for *op*, 49.4 to 50). Therefore, it was suggested that

the nonlinearity in this loop was most likely due to an oversized valve. This 6-inch valve was replaced by a 3-inch valve during the annual maintenance shutdown of the plant (in May, 2001). In order to confirm the result of the analysis, additional data were collected in July. The results of the 'post-maintenance' data analysis are shown in the right half of the Figure 3.19. For the new data set, $NGI = 0.0005$ (less than 0.001). These values indicate Gaussian linear system characteristics.

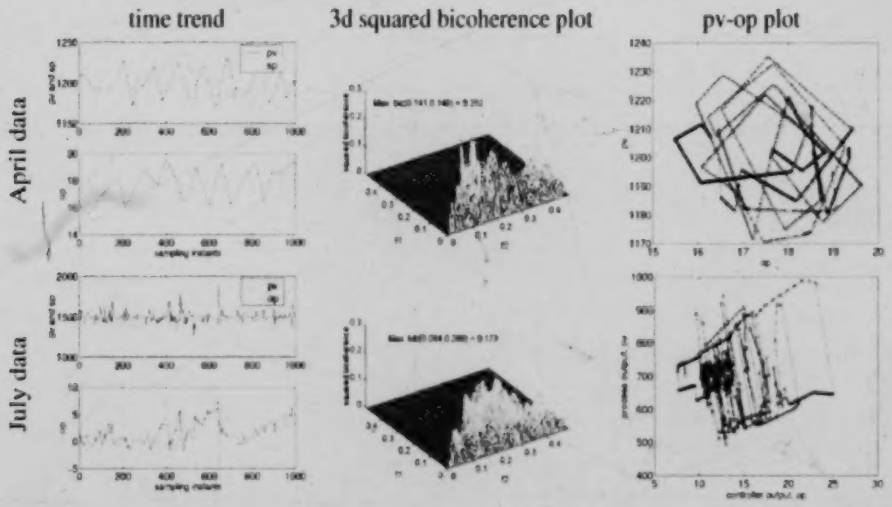


Figure 3.20: *Analysis of flow loop 2 data before (April - top) and after (July - bottom) the plant maintenance shutdown period.*

Corroded valve seat

This is also a flow control loop at the outlet of a pump located at the bottom of a distillation column. Analysis of the April, 2001 data for this loop revealed that the loop had nonlinearity problems with $NGI = 0.032$ and $NLI = 0.13$. The diagnostic plots are shown in the top panel of Figure 3.20. The test results correctly detected the presence of significant nonlinearity. The *pv-op* characteristic plot indicated a type of nonlinear characteristic in the process or the valve that had not been observed before. During the annual maintenance, plant instrument personnel noticed that the valve seat and the plug were severely corroded. Consequently, the valve was replaced. The results of the 'post-maintenance' analysis are

shown in the lower panel of the Figure 3.20. Now $NGI = 0.04$ and $NLI = 0.06$, indicating yet again the presence of a nonlinearity but in a substantially reduced form. The *pv-op* plot still shows unfamiliar patterns for unknown sources of nonlinearities. However, the overall controller performance of this loop has improved significantly to the point where additional analysis was deemed unnecessary.

3.8 Conclusions

Based on HOS theory, two new indices – the Non-Gaussianity Index (*NGI*) and the Non-Linearity Index (*NLI*) – have been developed to detect and quantify signal non-Gaussianity and nonlinearity. These indices together with specific patterns in the process output (*pv*) vs. the controller output (*op*) plot can be conveniently used to diagnose causes of poor control loop performance. The method has been successfully applied to many industrial data sets. Several representative examples of such studies have been presented here. In all cases, the results of the analysis were confirmed by plant engineers.

4

Modelling Control Valve Stiction

The presence of nonlinearities, e.g., stiction, deadband and backlash in a control valve limits the control loop performance. Stiction is the most common problem in spring-diaphragm type valves, which are widely used in the process industry. Though there have been many attempts to understand and model the stiction phenomenon, there is a lack of a proper model which can be understood and related directly to the practical situation as observed in real valves in process industry. This study focuses on the understanding, from real-world data, of the mechanism that causes stiction. It proposes a new data-driven model of stiction, which can be directly related to real valves. It also validates the simulation results generated using the proposed model with that from a physical model of the valve. Finally, valuable insights into stiction have been obtained from the describing function analysis of the newly proposed stiction model.

[1] *A full paper on this chapter has been accepted for publication in Control Engineering Practice, 2004 (Choudhury* et al., *2004b) and also a brief paper was presented at ADCHEM 2003 (Choudhury* et al., *2004a). All sections of this chapter were created entirely by the thesis author apart from the materials presented in sections 4.4 and 4.6, which are the results of a joint work with Nina Thornhill.*

4.1 Introduction

A typical chemical plant has hundreds or even thousands of control loops. Control performance is important to ensure high product quality and low cost of the product in such plants. The economic benefits resulting from performance assessment are difficult to quantify on a loop-by-loop basis because each problem loop contributes in a complicated way to the overall process performance. Finding and fixing problem loops throughout a plant reduces off-grade production, decreases product property variability, lowers operating costs and improves production rate (Paulonis and Cox, 2003). Even a 1% improvement either in energy efficiency or improved controller maintenance direction saves hundreds of millions of dollars for process industries (Desborough and Miller, 2002). Oscillatory variables are one of the main causes of poor performance of control loops. A key challenge is to find the root cause of distributed oscillations in chemical plants (Qin, 1998; Thornhill *et al.*, 2003*a*; Thornhill *et al.*, 2003*b*). The presence of oscillations in a control loop increases the variability of the process variables thereby resulting in inferior products, larger rejection rates, increased energy consumption, reduced average throughput and profitability. Oscillations can cause a valve to wear out much earlier than expected. Oscillations increase operating costs roughly in proportion to the deviation (Shinskey, 1990). Detection and diagnosis of the causes of oscillations in process operation are important. Clearly, a plant operating within product quality limits is more profitable than one that has to back away because of variations in the product (Martin *et al.*, 1991). Oscillatory feedback control loops commonly occur due to poor controller tuning, control valve stiction, poor process and control system design, and oscillatory disturbances (Bialkowski, 1992; Ender, 1993; Miao and Seborg, 1999). Bialkowski (1992) reported that about 30% of the loops are oscillatory due to control valve problems.

The only moving part in a control loop is the control valve. If the control valve contains nonlinearities, e.g., stiction, backlash, and deadband, the valve output may be oscillatory which in turn can cause oscillations in the process output. Among the many types of nonlinearities in control valves, stiction is the most common and one of the long-standing problems in the process industry. It hinders the achievement of good performance of control valves as well as control loops. Many studies (Horch, 2000; McMillan, 1995; Sharif and Grosvenor, 1998; Horch and Isaksson, 1998; Horch *et al.*, 2000; Aubrun *et al.*, 1995; Wallén, 1997; Taha *et al.*, 1996; Gerry and Ruel, 2001; Ruel, 2000; Armstrong-Hélouvry *et al.*, 1994) have been conducted out to define and detect static friction or stiction. However, a unique definition and description of the mechanism of stiction does not exist. The

present work addresses this issue as well as the modelling of valve friction or stiction. The parameters of a physical model, e.g., mass of the moving parts of the valve, spring constants, and forces, are not explicitly known. These parameters need to be tuned properly to produce the desired response of a valve. The effect of changes in these parameters are also unknown. Working with such a physical model is therefore often time-consuming and cumbersome for simulation purposes. Also, in industrial practice, stiction and other related problems are identified in terms of the percentage of the valve travel or span of the valve input signal. The relationship between the magnitudes of the parameters of a physical model and deadband, backlash or stiction (expressed as a % of the span of the input signal) is not simple. The purpose of this study is to develop an empirical, data-driven model of stiction that is useful for simulation and diagnosis of oscillation in chemical processes. The main contributions of this chapter are:

- An attempt to clarify the confusion prevalent in the control literature and in the control community regarding the misunderstanding of stiction and the terms associated with it.

- A new formal definition of stiction has been proposed using parameters similar to those used in the American National Standard Institution's (ANSI) formal definition of backlash, hysteresis, and deadband. The key feature of these definitions is that they focus on the input-output behaviour of such elements. The proposed definition is also cast in terms of the input-output behaviour.

- A new two-parameter data-driven model of stiction has been developed and validated with a mechanistic model of stiction and also with data obtained from industrial control valves suffering from stiction. The data-driven model is capable of handling stochastic inputs and can be be used to perform simulation of stiction in MATLAB®'s Simulink environment in the studies of stiction relevant control loop problems. For example, stiction model can be used to create a root cause for originating oscillation in a plant-wide oscillation study.

- A describing function analysis of the newly proposed stiction model reveals valuable insights on stiction behaviour. For example, pure deadband or backlash cannot produce limit cycles in the presence of a PI controller unless there is an integrator in the plant under a closed loop feedback configuration.

This chapter has been organized as follows. First, a thorough discussion of the terms related to valve nonlinearity are presented, followed by the proposal of a new formal

definition of stiction. Some practical examples of valve stiction are provided to gain insights into stiction from real-world data. The results of a mechanistic model of stiction are used to validate the corresponding subsequent results of the data-driven stiction model. Finally, a describing function analysis of the newly proposed stiction model is presented.

4.2 What is Stiction?

Terms such as deadband, backlash and hysteresis are often misused and wrongly used in describing valve problems. For example, a deadband in a valve is commonly referred to as *backlash* or *hysteresis*. Therefore, before proceeding to the definition of stiction, these terms must be defined for a better understanding of the stiction mechanism and a more formal definition of stiction.

4.2.1 Definition of terms relating to valve nonlinearity

This section reviews the American National Standard Institution's (ANSI) formal definition of terms related to stiction. The aim is to differentiate clearly between the key concepts that underlie the ensuing discussion of friction in control valves. These definitions can also be found in (EnTech, 1998; Fisher-Rosemount, 1999), which also refer to ANSI. ANSI (ISA-S51.1-1979, Process instrumentation Terminology) defines the above terms as follows:

- Backlash: *"In process instrumentation, it is a relative movement between interacting mechanical parts, resulting from looseness, when the motion is reversed"*.
- Hysteresis: *"Hysteresis is that property of the element evidenced by the dependence of the value of the output, for a given excursion of the input, upon the history of prior excursions and the direction of the current traverse."*
  - *"It is usually determined by subtracting the value of deadband from the maximum measured separation between upscale going and downscale going indications of the measured variable (during a full range traverse, unless otherwise specified) after transients have decayed."* Figure 4.1(a) and (c) illustrates the concept.
  - *"Some reversal of output may be expected for any small reversal of input. This distinguishes hysteresis from deadband."*

- Dead band: *"In process instrumentation, it is the range through which an input signal may be varied, upon reversal of direction, without initiating an observable change in output signal."*
  - *"There are separate and distinct input-output relationships for increasing and decreasing signals (See Figure 4.1(b))."*
  - *"Deadband produces phase lag between input and output."*
  - *"Deadband is usually expressed in percent of span."*

  Deadband and hysteresis may be present simultaneously. In such an event, the characteristics in the lower left panel of Figure 4.1 would be observed.

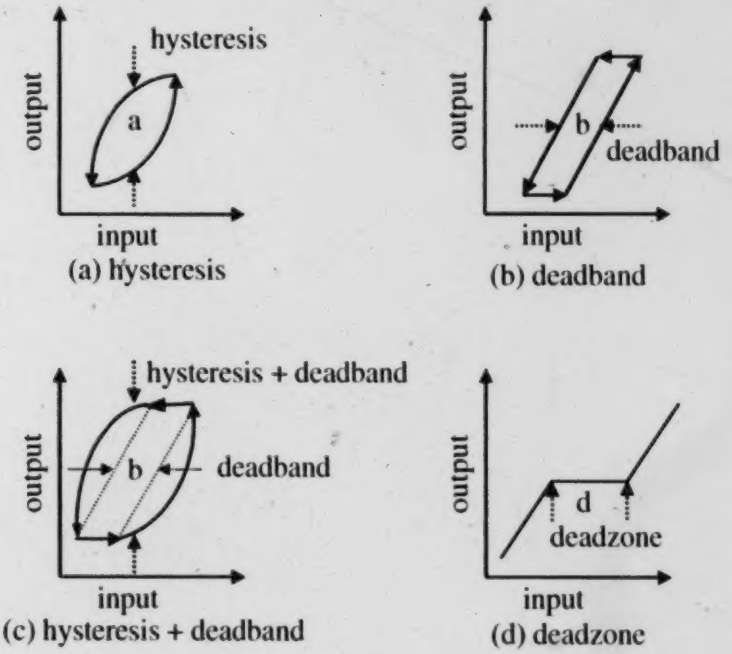


Figure 4.1: *Input-Output behaviour of hysteresis, deadband, and deadzone (redrawn from ANSI/ISA-S51.1-1979).*

- Dead Zone: *"It is a predetermined range of input through which the output remains unchanged, irrespective of the direction of change of the input signal." (There is another definition of deadzone for multi-position controller. It is a zone of input for which no value of output exists. It is usually intentional and adjustable)*

- *"There is but one input-output relationship (See Figure 4.1(d))."*
- *"Dead zone produces no phase lag between input and output."*

The above definitions show that the term "backlash" applies specifically to the slack or looseness of the mechanical part when the motion changes direction. Therefore, in control valves it may add deadband effects only if there is some slack in rack-and-pinion type actuators (Fisher-Rosemount, 1999) or loose connections in a rotary valve shaft. ANSI (ISA-S51.1-1979) definitions and Figure 4.1 show that hysteresis and deadband are distinct effects. Deadband is quantified in terms of input signal span (i.e., on the x-axis), while hysteresis refers to a separation in the measured (output) response (i.e., on the y-axis).

4.2.2 Discussion of the term "Stiction"

Various people and organizations have defined stiction in various ways. Several definitions are reproduced below:

- According to the Instrument Society of America (ISA) (ISA Subcommittee SP75.05, 1979), "*stiction is the resistance to the start of motion, usually measured as the difference between the driving values required to overcome static friction upscale and downscale*". The definition was first proposed in 1963 in American National Standard C85.1-1963,"Terminology for Automatic Control" and has not been updated. This definition was adopted in ISA 1979 Handbook (ISA Subcommittee SP75.05, 1979) and has remained the same in the revised 1993 edition.

- According to Entech (1998), "*stiction is a tendency to stick-slip due to high static friction. The phenomenon causes a limited resolution of the resulting control valve motion. ISA terminology has not settled on a suitable term yet. Stick-slip is the tendency of a control valve to stick while at rest, and to suddenly slip after force has been applied*".

- According to (Horch, 2000), "*The control valve is stuck in a certain position due to high static friction. The (integrating) controller then increases the set point to the valve until the static friction can be overcome. Then the valve breaks off and moves to a new position (slip phase) where it sticks again. The new position is usually on the other side of the desired set point such that the process starts in the opposite direction again*". This is the extreme case of stiction. On the contrary, once the valve

overcomes stiction, it might travel smoothly for some time and then stick again when the velocity of the valve is close to zero.

- In a recent paper, Ruel (2000) reported "*stiction as a combination of the words stick and friction, created to emphasize the difference between static and dynamic friction. Stiction exists when the static (starting) friction exceeds the dynamic (moving) friction inside the valve. Stiction describes the valve's stem (or shaft) sticking when small changes are attempted. Friction of a moving object is less than when it is stationary. Stiction can keep the stem from moving for small control input changes, and then the stem moves when there is enough force to free it. The result of stiction is that the force required to get the stem to move is more than is required to go to the desired stem position. In presence of stiction, the movement is jumpy*".

  This definition resembles stiction as measured online in process industries — putting the control loop in manual and then increasing the valve input in small increments until there is a noticeable change in the process variable.

- In (Olsson, 1996), stiction is defined as "*short for static friction as opposed to dynamic friction. It describes the friction force at rest. Static friction counteracts external forces below a certain level and thus keeps an object from moving*".

The above discussion reveals the lack of a formal and general definition of stiction and the mechanism(s) that causes it. All of the above definitions agree that stiction is the static friction that keeps an object from moving and when the external force overcomes the static friction the object starts moving. However, these definitions disagree in the way stiction is measured and how it can be modelled. Also, there is no clear description of what happens at the moment when the valve just overcomes the static friction. Several modelling approaches described this phenomenon using a Stribeck effect model (Olsson, 1996).

4.2.3 A proposal for a definition of stiction

The motivation for a new definition of stiction is to capture the descriptions cited earlier within a definition that explains the behaviour of a valve with stiction in terms of its input-output behaviour, as is done in the ANSI definitions for backlash, hysteresis, and deadband. The new definition of stiction as proposed by the author is based on careful investigation of real process data. It is observed that the phase plot of the input-output behaviour of a valve "suffering from stiction" can be described as shown in Figure 4.2. It consists of four components: deadband, stickband, slip-jump and the moving phase. When the valve

comes to rest or changes direction at point A in Figure 4.2, the valve sticks. After the controller output overcomes the deadband (AB) and the stickband (BC) of the valve, the valve jumps to a new position (point D) and continues to move. Due to very low or zero velocity, the valve may stick again between points D and E in Figure 4.2 while travelling in the same direction (EnTech, 1998). In such a case, the magnitude of deadband is zero and only stickband is present. This can be overcome only if the controller output signal is larger than the stickband. The latter is uncommon in industrial practice.

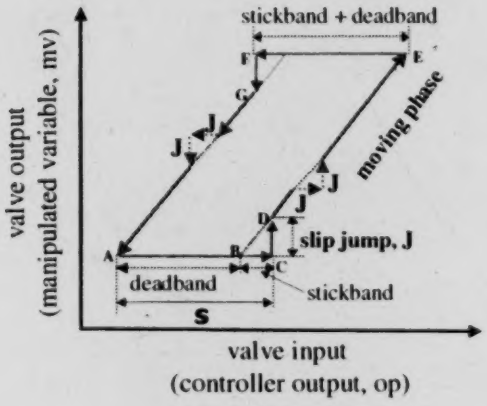


Figure 4.2: *Typical input-output behaviour of a sticky valve.*

The deadband and stickband represent the behaviour of the valve when it is not moving, even though the input to the valve keeps changing. Slip-jump represents the abrupt release of potential energy stored in the actuator chambers due to high static friction in the form of kinetic energy as the valve starts to move. The magnitude of the slip-jump is critical in determining the limit cyclic behaviour introduced by stiction (McMillan, 1995; Piipponen, 1996). Once the valve slips, it continues to move until it sticks again (point E in Figure 4.2). In this moving phase, dynamic friction may be much lower than the static friction. As depicted in Figure 4.2, this section has proposed a rigorous description of the effects of friction in a control valve. Therefore, we propose the following new definition of stiction:

"Stiction is a property of an element whose smooth movement in response to a varying input is preceded by a stickband and an abrupt jump termed the slip-jump. Slip-jump is

expressed as a percentage of the output span. Its origin in a mechanical system is static friction, which exceeds the friction during smooth movement".

This definition has been exploited in the next and subsequent sections for the evaluation of practical examples and for modelling of a control valve suffering from stiction in a feedback control configuration.

4.3 Practical Examples of Valve Stiction

The objective of this section is to explore effects of stiction through the investigation of data from industrial control loops. The observed effects reinforce the need for a rigorous definition of stiction. This section analyzes four data sets. The first set is from a power plant, the second and third are from a petroleum refinery and the fourth is from a furnace. To preserve the confidentiality of the plants, all data are scaled and reported as mean-centered with unit variance. The notations followed by industry are used here in order to enhance the readability of this work. For example, *pv* is used to denote the process variable or controlled variable. Similarly, *op* is used to denote the controller output, *mv* to denote valve output or valve position, and *sp* to denote set point.

- Loop 1 is a level control loop that controls the level of condensate in the outlet of a turbine by manipulating the flow rate of the liquid condensate. In total, 8640 samples for each tag were collected at a sampling rate of 5 seconds. Figure 4.3 shows a portion of the time domain data. The left panel shows time trends for level (*pv*), the controller output (*op*) which is also the valve demand, and valve position (*mv*) which can be taken to be the same as the condensate flow rate. The plots in the right panel show the characteristics *pv-op* and *mv-op* plots. The bottom plot clearly indicates both the stickband plus deadband and the slip-jump effects. The slip-jump is large and visible from the bottom plot especially, when the valve is moving in a downward direction. It is marked as 'A' in the figure. It is evident from this figure that the valve output (*mv*) can never reach the valve demand (*op*). This kind of stiction is termed as 'undershoot case' of valve stiction in this study. The *pv-op* plot does not show the jump behaviour clearly. The slip-jump is very difficult to observe in the $pv - op$ plot because the process dynamics (i.e., the transfer function between *mv* and *pv*) destroy the pattern. This loop shows one possible type of stiction phenomenon clearly. The stiction model developed later in the paper based on the control signal (*op*) is able to imitate this kind of behaviour.

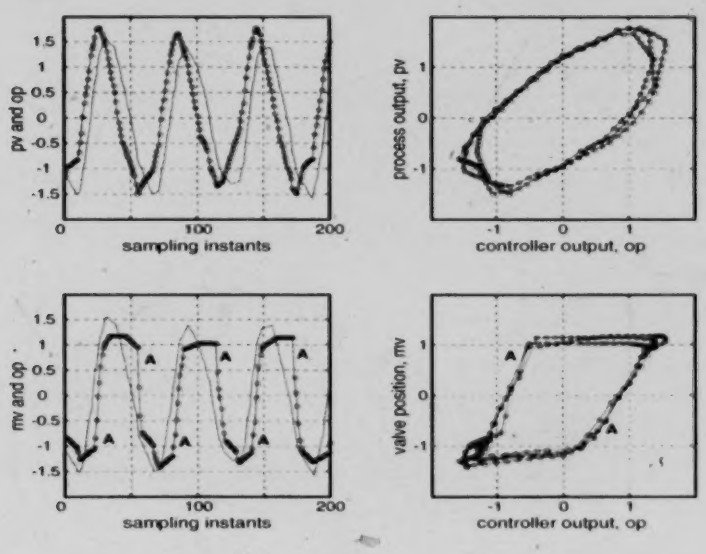


Figure 4.3: *Flow control cascaded to level control in an industrial setting, the line with circles is pv and mv, the thin line is op.*

- Loop 2 is a liquid flow slave loop of a cascade control loop. The data was collected at a sampling rate of 10 s and the data length for each tag was 1000 samples. The left plot of Figure 4.4 shows the time trend of *pv* and *op*. A closer examination of this figure reveals that the *pv* (flow rate) is constant for a given period of time, though the *op* changes over that period. This is the period during which the valve was stuck. Once the valve overcomes deadband plus stickband, the *pv* changes very quickly (denoted as 'A' in the figure) and moves to a new position where the valve sticks again. It is also evident that sometimes the *pv* overshoots the *op*, and sometimes it undershoots. The *pv*-*op* plot has two distinct parts – the lower part and the upper part extended to the right. The lower part corresponds to the overshoot case of stiction, i.e, it represents an extremely sticky valve. The upper part corresponds to the undershoot case of stiction. These two cases have been separately modelled in the data driven stiction model. This example constitutes a mixture of undershoot and overshoot cases of stiction. The terminologies regarding different cases of stiction will be clarified in Section 4.5.

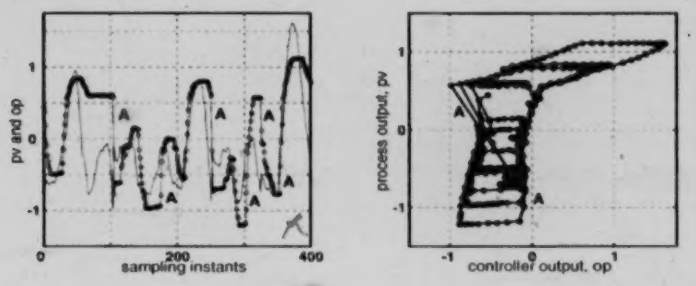


Figure 4.4: *Data from a flow loop in a refinery, time trend of pv and op (left) - the line with circles is pv and the thin line is op, and the pv-op plot (right).*

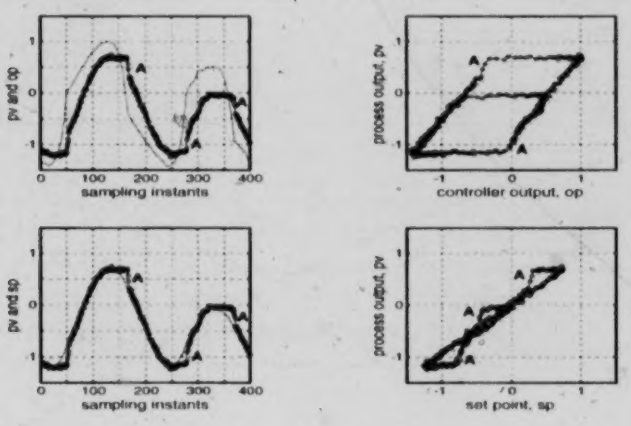


Figure 4.5: *Data from a flow loop in a refinery, time trend of pv and op (top left) - the line with circles is pv and the thin line is op, the pv-op plot(top right), time trend of pv and sp (bottom left), line with circles is pv and thin line is sp, and the pv-sp plot(bottom right).*

- Loop 3 is a slave flow loop cascaded with a master level control loop. A sampling rate of 6 s was used for the collection of data and a total of 1000 samples for each tag were collected. The top panel of Figure 4.5 shows the presence of stiction with a clear indication of stickband plus deadband and the slip-jump phase. The slip-jump appears as the control valve just overcomes stiction (denoted as point 'A' in Figure 4.5). This slip-jump is not very clear in the *pv-op* plot of the closed loop data (top right plot) because both *pv* and *op* jump together due to the probable presence of a proportional only controller. However, it shows the presence of deadband plus stickband clearly. Sometimes it is best to look at the *pv-sp* plot when it is a cascaded loop and the slave loop is operating under proportional control only. The bottom panel of Figure 4.5 shows the time trend and phase plot of *sp* and *pv*, where the slip-jump behaviour is clearly visible. This example represents a case of pure stick-slip or stiction with no offset.

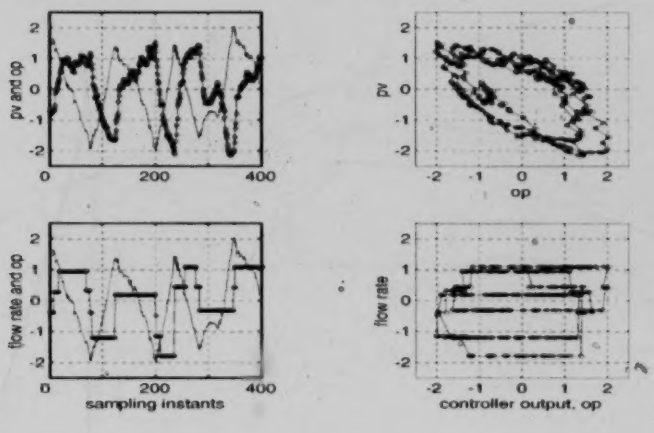


Figure 4.6: *Industrial dryer temperature control loop data, lines with circles are pv and flowrate, the thin line is op (the bottom left panel).*

- Loop 4 is a temperature control loop on a furnace feed dryer system at the Tech-Cominco mine in Trail, British Columbia, Canada. The temperature of the dryer combustion chamber is controlled by manipulating the flow rate of natural gas to the combustion chamber. A total 1440 samples for each tag were collected at a sampling rate of 1 min. The top plot of the left panel of the Figure 4.6 shows time trends of

temperature (*pv*) and controller output (*op*). It shows clear oscillations both in the controlled variable (*pv*) and the controller output. The presence of distinct loops is observed in the characteristic *pv-op* plot (see Figure 4.6 top right). For this loop, there was a flow indicator close to this valve and this indicator data were available. In the bottom figure this flow rate is plotted versus *op*. The flow rate data appear to be quantized but the presence of stiction in this control valve was confirmed by the plant engineer. The two bottom plots clearly show the stickband and the slip-jump of the valve. Note that the moving phase of the valve is almost absent in this example. After the valve overcomes stiction, it jumps to the new position and sticks again.

4.4 A Physical Model of Valve Friction

4.4.1 Model formulation

The purpose of this section is to understand the physics of valve friction and reproduce the behaviour seen in real plant data. A cross-sectional diagram of a typical pneumatic control valve is shown in Figure 4.7. For such a pneumatic sliding stem valve, the force balance equation based on Newton's second law can be written as:

$$M\frac{d^2x}{dt^2} = \sum Forces = F_a + F_r + F_f + F_p + F_i \tag{4.1}$$

where M is the mass of the moving parts, x is the relative stem position, $F_a = Au$ is the force applied by pneumatic actuator where A is the area of the diaphragm and u is the actuator air pressure or the valve input signal, $F_r = -kx$ is the spring force where k is the spring constant, $F_p = -A_p\Delta P$ is the force due to fluid pressure drop where A_p is the plug unbalance area and ΔP is the fluid pressure drop across the valve, F_i is the extra force required to force the valve to be into the seat and F_f is the friction force (Fitzgerald, 1995; Kayihan and Doyle III, 2000; Whalen, 1983). Following Kayihan and Doyel III, F_i and F_p will be assumed to be zero because of their negligible contribution in the model.

The friction model is from (Karnopp, 1985; Olsson, 1996) and was also used by Horch and Isaksson (1998). It includes static and moving friction. The expression for the moving friction is in the first line of Equation 4.2 and comprises a velocity independent term F_c known as Coulomb friction and a viscous friction term vF_v that depends linearly upon the

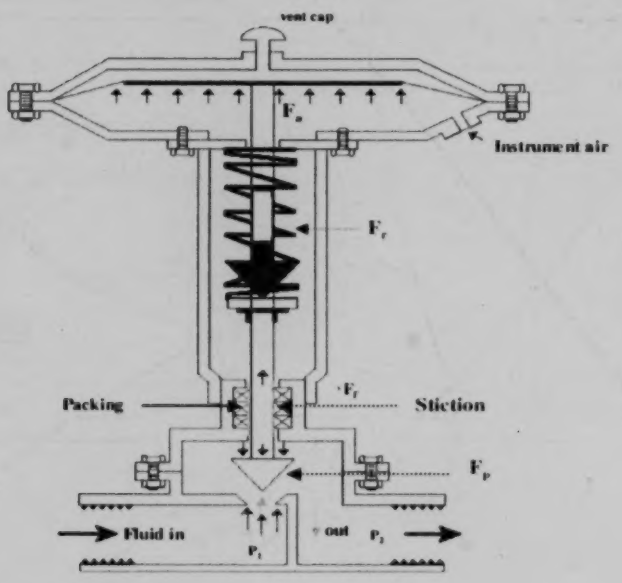


Figure 4.7: *A cross-sectional diagram of a pneumatic control valve.*

velocity. Both act in opposition to the velocity, as shown by the negative signs.

$$F_f = \begin{cases} -F_c\, sgn(v) - vF_v & \text{if} \quad v \neq 0 \\ -(F_a + F_r) & \text{if} \quad v = 0 \text{ and } |F_a + F_r| \leq F_s \\ -F_s\, sgn(F_a + F_r) & \text{if} \quad v = 0 \text{ and } |F_a + F_r| > F_s \end{cases} \tag{4.2}$$

The second line in Equation 4.2 is the case when the valve is stuck. F_s is the maximum static friction. The velocity of the stuck valve is zero and not changing, therefore the acceleration is zero also. Thus, the right-hand side of Newton's law is zero, such that $F_f = -(F_a + F_r)$. The third line of the model represents the situation at the instant of breakaway. At that instant the sum of forces is $(F_a + F_r) - F_s\, sgn(F_a + F_r)$, which is not zero if $|F_a + F_r| > F_s$. Therefore, the acceleration becomes non-zero and the valve starts to move.

A disadvantage of a physical model of a control valve is that it requires several parameters (M, F_s, F_c, F_v) to be known. The mass M and typical friction forces (F_s, F_c, F_v) depend upon the design of the valve. Kayihan and Doyle III (2000) used manufacturer's values suggested by Fitzgerald (1995) and similar values have been chosen here apart from a slightly increased value of F_s and a smaller value for F_c in order to make the demonstration of the slip-jump more obvious (see Table 4.1). Figure 4.8 shows the friction

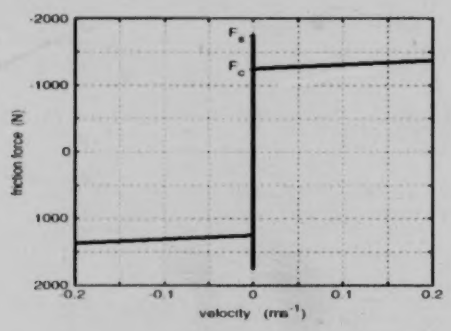


Figure 4.8: *Friction characteristic plot.*

force characteristic in which the magnitude of the moving friction is smaller than that of the static friction. The friction force opposes velocity (see Equation 4.2) thus the force is negative when the velocity is positive.

Table 4.1: Nominal values used for physical valve simulation

Parameters	Kayihan and Doyle III, 2000	Nominal case
M	3 lb (1.36 kg)	1.36 kg
F_s	384 lbf (1708 N)	1750 N
F_c	320 lbf (1423 N)	1250 N
F_v	3.5 $lbf.s.in^{-1}$ (612 $N.s.m^{-1}$)	612 $N.s.m^{-1}$
spring constant, k	300 $lbf.in^{-1}$ (52500 $N.m^{-1}$)	52500 $N.m^{-1}$
diaphragm area, A	100 in^2 (0.0645 m^2)	0.0645 m^2
calibration factor, k/A	-	807692 $Pa.m^{-1}$
air pressure	10 *psi* (68950 Pa)	68950 Pa

The calibration factor of Table 4.1 is introduced because the required stem position x_r is the input to the simulation. In the absence of stiction effects the valve moving parts come to rest when the force due to air pressure on the diaphragm is balanced by the spring force. Thus $Au = kx$ and the calibration factor relating air pressure u to x_r is k/A. The consequences of miscalibration are discussed below.

4.4.2 Valve simulation

The purpose of simulation of the valve was to determine the influence of the three friction terms in the model. The nonlinearity in the model is able to induce limit cycle oscillations in a feedback control loop. The aim is to understand the contribution of each friction term to the character and shape of the limit cycles.

Open Loop response:

Figure 4.9 shows the valve position when the valve model is driven by a sinusoidal variation in op in the absence of the controller. The left-hand column shows the time trends while the right hand panels are plots of valve demand (op) versus valve position (mv). Several cases are simulated using the parameters shown in Table 4.2. The "linear" values are those suggested by Kayihan and Doyle III for the best case of a smart valve with Teflon packing requiring air pressure of about 0.1 *psi* (689 *Pa*) to start it moving.

Table 4.2: Friction values used in simulation of physical valve model

Parameters	linear	pure deadband	stiction (undershoot) (open loop)	stiction (undershoot) (closed loop)	stiction (no offset)
$F_s\,(N)$	45	1250	2250	1000	1750
$F_c\,(N)$	45	1250	1250	400	0
$F_v\,(N.s.m^{-1})$	612	612	612	612	612

In the first row of Figure 4.9, the Coulomb friction F_c and static friction F_s are small and linear viscous friction dominates. The input and output are almost in phase in the first row of Figure 4.9 because the sinusoidal input is of low frequency compared to the bandwidth of the valve model and is on the part of the frequency response function where input and output are in phase.

Valve deadband is due to the presence of Coulomb friction F_c, a constant friction which acts in the opposite direction to the velocity. In the deadband simulation case the static friction is the same as the Coulomb friction, $F_s = F_c$. The deadband arises because, on changing direction, the valve remains stationary until the net applied force is large enough to overcome F_c. The deadband becomes larger if F_c is larger.

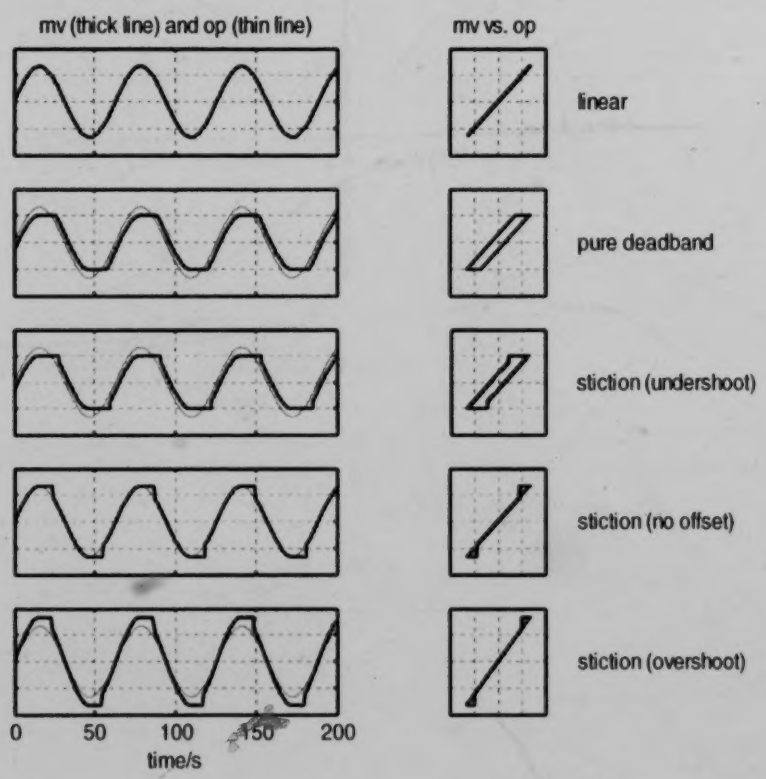


Figure 4.9: *Open loop response of mechanistic model. The amplitude of the sinusoidal input is 10 cm in each case.*

A valve with high initial static friction such that $F_s > F_c$ exhibits a jumping behaviour that is different from a deadband, although both behaviors may be present simultaneously. When the valve starts to move, the friction force decreases abruptly from F_s to F_c. There is therefore a discontinuity in the model on the right hand side of Newton's second law and a large increase in acceleration of the valve moving parts. The initial velocity is therefore higher than in the $F_s = F_c$ case, leading to the jump behaviour observed in the third row of Figure 4.9. If the Coulomb friction F_c is absent, then the deadband is absent and the slip-jump allows the *mv* to catch up with the *op* (fourth row). If the valve is miscalibrated, then swings in the valve position (*mv*) are larger than swings in the demanded position (*op*). In that case the gradient of the *op-mv* plot is greater than unity during the moving phase. The

bottom row of Figure 4.9 shows the case when the calibration factor is too large by 25%. A slip-jump was also used in this simulation.

Closed loop dynamics:

For assessment of closed loop behaviour, the valve output drives a first order plus dead time process $G(s)$ and receives its *op* reference input from a PI controller $C(s)$, where:

$$G(s) = \frac{3e^{-10s}}{10s+1} \qquad C(s) = 0.2\left(\frac{10s+1}{10s}\right) \tag{4.3}$$

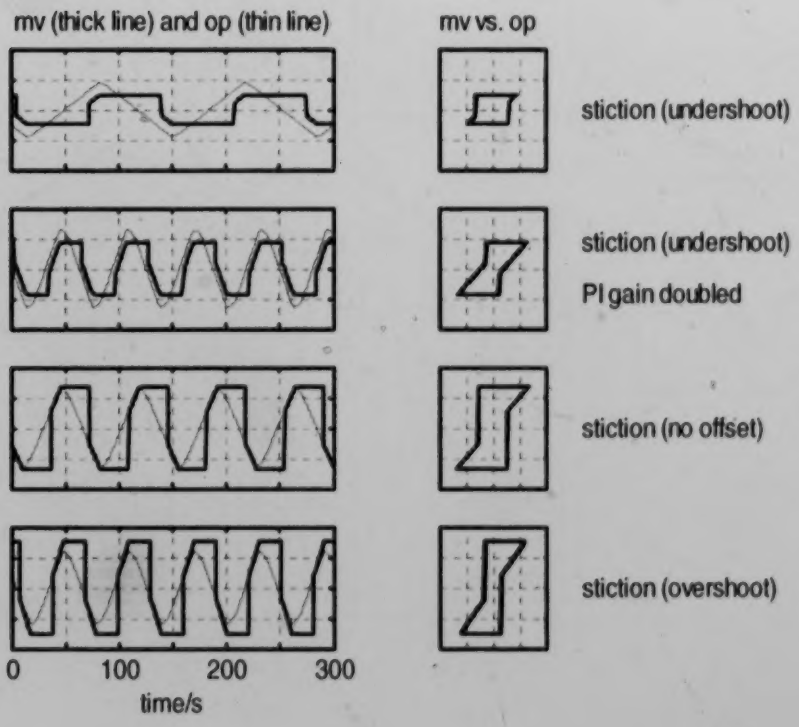


Figure 4.10: *Closed loop response of mechanistic model.*

Figure 4.10 shows the limit cycles induced in this control loop by the valve together with the plots of valve position (*mv*) versus valve demand (*op*). The limit cycles were present even though the set point to the loop was zero. That is, they were internally generated and sustained by the loop in the absence of any external setpoint excitation.

There was no limit cycle in the linear case dominated by viscous friction or in the case with deadband only when $F_s = F_c$. It is known that deadband alone cannot induce a limit cycle unless the process $G(s)$ has integrating dynamics, as will be discussed further in Section 4.5.3.

The presence of stiction ($F_s > F_c$) induces a limit cycle with a characteristic triangular shape in the controller output. Cycling occurs because an offset exists between the set point and the output of the control loop while the valve is stuck which is integrated by the PI controller to form a ramp. By the time the valve finally moves in response to the controller *op* signal the actuator force has grown quite large and the valve moves quickly to a new position where it then sticks again. Thus, a self limiting cycle is set up in the control loop.

If stiction and deadband are both present, then the period of the limit cycle oscillation can become very long. The combination $F_s = 1750\,N$ and $F_c = 1250\,N$ gave a period of 300s while the combination $F_s = 1000\,N$ and $F_c = 400\,N$ had a period of about 140s (top row, Figure 4.10). In both cases, the period is much longer than the time constant of the controlled process or its cross-over frequency. The period of oscillation can also be influenced by altering the controller gain. If the gain is increased the linear ramps of the controller output signal are steeper, the actuator force moves through the deadband more quickly and the period of the limit cycle becomes shorter (second row, Figure 4.10). The technique of changing the controller gain is used by industrial control engineers to test the hypothesis of a limit cycle induced by valve non-linearity while the plant is still running in closed loop.

In the pure stick-slip or stiction with no offset case shown in the third row of Figure 4.10 the Coulomb friction is negligible and the oscillation period is shorter because there is no deadband. The bottom row in Figure 4.10 shows that miscalibration causes an overshoot in closed loop.

4.5 Data Driven Model of Valve Stiction

The proposed data-driven model has parameters that can be related directly to plant data and it produces the same behaviour as the physical model. The model needs only an input signal and the specification of deadband plus stickband and slip-jump. It overcomes the main disadvantages of physical modelling of a control valve, namely it requires the knowledge of the mass of the moving parts of the actuator, spring constant, and the friction forces. The effect of the change of these parameters can not easily be determined analytically

because the relationship between the values of the parameters and the observation of the deadband/stickband as a percentage of valve travel is not straightforward. In a data-driven model, the parameters are easy to choose and the effects of these parameter changes are simple to realize.

4.5.1 Model formulation

The valve sticks only when it is at rest or changing its direction. When the valve changes its direction it comes to rest momentarily. Once the valve overcomes stiction, it starts moving and may keep moving for sometime, depending on how much stiction is present in the valve. In this moving phase, it suffers only dynamic friction, which may be smaller than static friction. It continues to do so until its velocity is again very close to zero or it changes its direction.

In the process industry, stiction is generally measured as a % of the valve travel or the span of the control signal (Gerry and Ruel, 2001). For example, a 2 % stiction means that when the valve gets stuck, it will start moving only after the cumulative change of its control signal is greater than or equal to 2%. If the range of the control signal is 4 to 20 mA, then a 2% stiction means that a change of the control signal less than 0.32 mA in magnitude will not be able to move the valve.

In the modelling approach described herein, the control signal has been translated to the percentage of valve travel with the help of a linear look-up table. If the control signal is noisy, then a filter, e.g. exponentially weighted moving average filter (EWMA), can be used to filter the noise. The model consists of two parameters – namely the size of deadband plus stickband S (specified in the input axis) and slip-jump J (specified on the output axis). Note that the term 'S' contains both the deadband and stickband. Figure 4.11 summarizes the model algorithm, which can be described as:

- First, the controller output (mA) is converted to valve travel % using a look-up table.
- If the transformed controller output signal (%) is less then 0 or more than 100, the valve is saturated (i.e., fully close or fully open).
- If the signal is within the 0 to 100% range, the algorithm calculates the slope of the controller output signal.
- Next, the change of the direction of the slope of the input signal is taken into consideration. If the 'sign' of the slope changes or remains zero for two consecutive

instants, the valve is assumed to be stuck and does not move. The 'sign' function of the slope gives the following:

- If the slope of input signal is positive, the sign(slope) returns '+1'.
- If the slope of input signal is negative, the sign(slope) returns '-1'.
- If the slope of input signal is zero, the sign(slope) returns '0'.

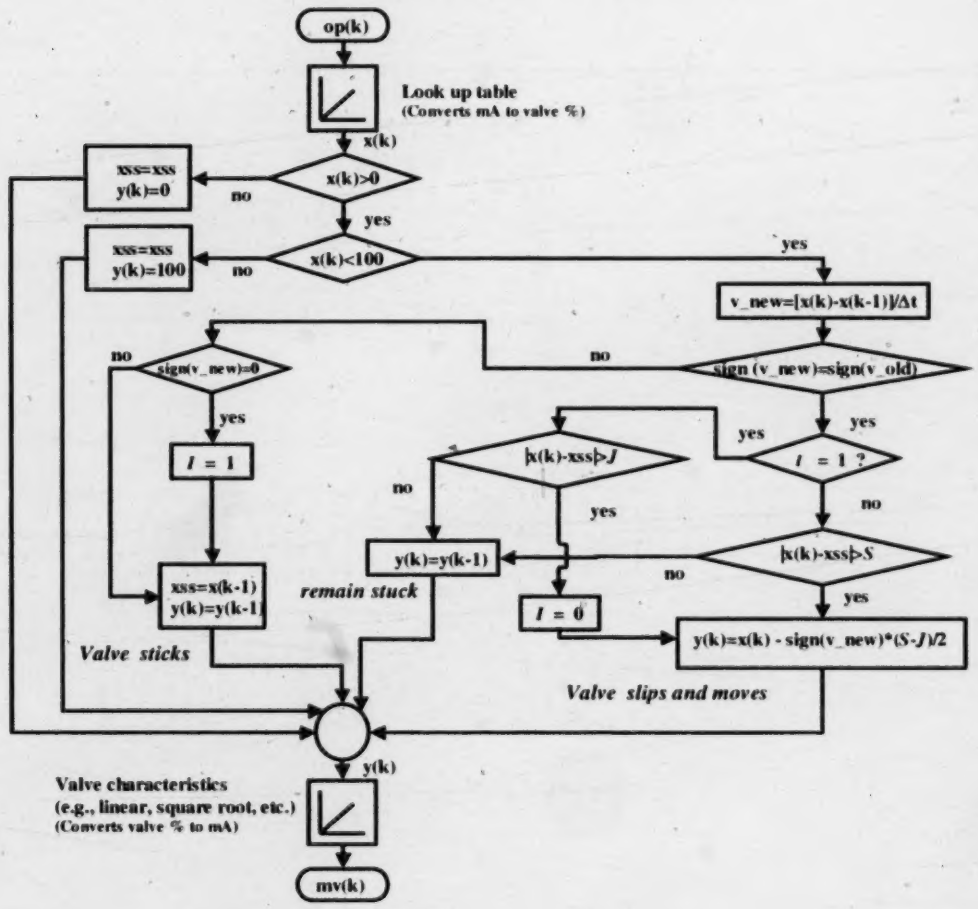


Figure 4.11: *Signal and logic flow chart of the data-driven stiction model.*

Therefore, when sign(slope) changes from '+1' to '-1' or vice versa, this means that the direction of the input signal has been changed and the valve is in the beginning of its stick position (points A and E in Figure 4.2). The algorithm detects stick position

of the valve at this point. Now, the valve may stick again while travelling in the same direction (opening or closing direction) only if the input signal to the valve does not change or remains constant for two consecutive instants, which is uncommon in practice. For this situation, the sign(slope) changes to '0' from '+1' or '-1' and vice versa. The algorithm again detects the stick position of the valve in the moving phase and this stuck condition is denoted with the indicator variable $I = 1$. The value of the input signal when the valve gets stuck is denoted as *xss*. This value of *xss* is kept constant and does not change until the valve gets stuck again. The cumulative change of input signal to the model is calculated from the deviation of the input signal from *xss*.

- For the case where the input signal changes direction (i.e., the sign(slope) changes from '+1' to '-1' or vice versa), if the cumulative change of the input signal is more than the amount of the deadband plus stickband (S), then the valve slips and starts moving.

- For the case when the input signal does not change direction (i.e., the sign(slope) changes from '+1' or '-1' to zero, or vice versa), if the cumulative change of the input signal is more than the amount of the stickband (J), then the valve slips and starts moving. Note that this takes care of the case when valve sticks again while traveling in the same direction (EnTech, 1998; Kano *et al.*, 2004).

- The output is calculated using the equation:

$$output = input - sign(slope) * (S - J)/2 \tag{4.4}$$

and depends on the type of stiction present in the valve. It can be described as follows:

  - Deadband: If $J = 0$, then it represents pure deadband case without any slip-jump.
  - Stiction (undershoot): If $J < S$, then the valve output can never reach the valve input. There is always some offset. This represents the undershoot case of stiction.
  - Stiction (no offset): If $J = S$, the algorithm produces pure stick-slip behaviour. There is no offset between the input and output. Once the valve overcomes stiction, valve output tracks the valve input exactly. This is the well-known **"stick-slip case"**.

  - Stiction (overshoot): If $J > S$, the valve output overshoots the valve input due to excessive stiction. This is termed as overshoot case of stiction.

  Recall that J is an output (y-axis) quantity. Also, the magnitude of the slope between input and output is 1.

- The parameter J signifies the slip-jump start of the control valve immediately after it overcomes the deadband plus stickband. It accounts for the offset between the valve input and output signals.

- Finally, the output is converted back to an mA signal using a look-up table based on characteristics of the valve such as linear, equal percentage or square root, and the new valve position is reported.

4.5.2 Open loop response of the model under a sinusoidal input

Figure 4.12 shows the open loop behaviour of the new data-driven stiction model in the presence of various types of stiction. Plots in the left panel show the time trend of the valve input *op* (thin solid line) and the output *mv* (thick solid line). The right panel shows the input-output behaviour of the valve on a X-Y plot.

- The first row shows the case of a linear valve without stiction.

- The second row corresponds to pure deadband without any slip jump, i.e., $J = 0$. Note that for this case, the magnitude of stickband is zero and deadband itself equals 'S'.

- The third row shows the undershoot case of a sticky valve where $J < S$. This case is illustrated in the first and second examples of industrial control loops. In this case, the valve output can never reach the valve input. There is always some offset.

- The fourth row represents pure stick-slip behaviour. There is no offset between the input and output. Once the valve overcomes stiction, valve output tracks the valve input accurately.

- In the fifth row, the valve output overshoots the desired set position or the valve input due to excessive stiction. This is termed "overshoot case of stiction".

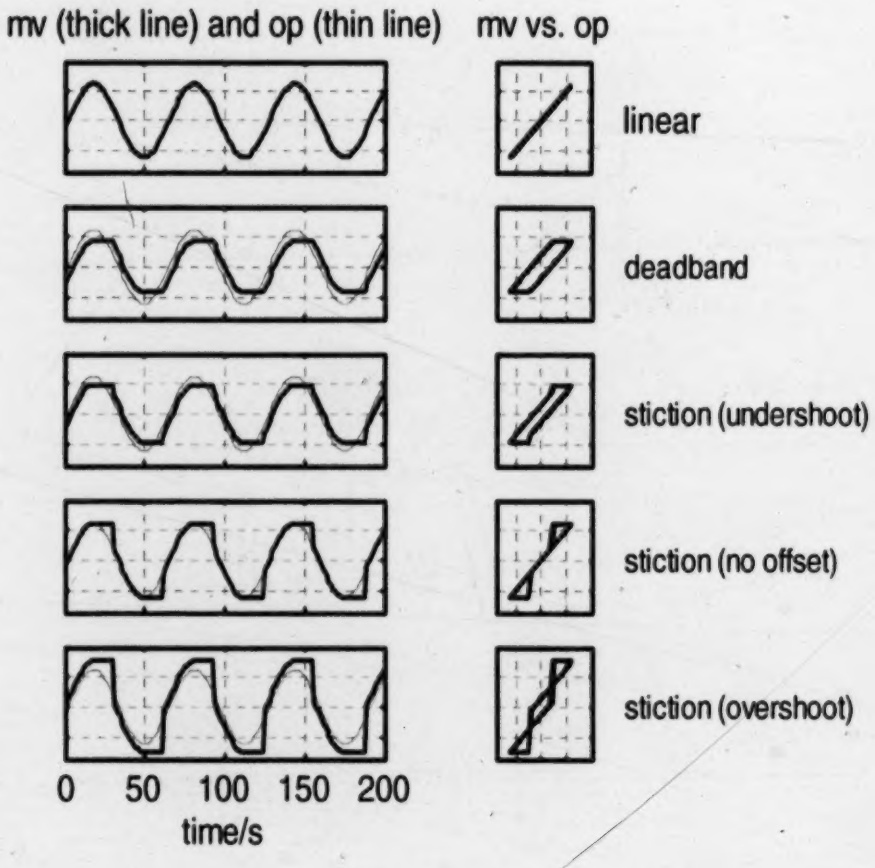


Figure 4.12: *Open loop simulation results of the data-driven stiction model.*

In reality, a composite of these stiction phenomena may be observed. Although this model is not directly based on the dynamics of the valve, the strength of the model is that it is very simple to use for the purpose of simulation and can quantify stiction as a percentage of valve travel or span of input signal. Also, the parameters used in this model are easy to understand, realize and relate to real stiction behaviour. Though this is an empirical model and not based on physics, it is observed that this model can correctly reproduce the behaviour of the physics-based stiction model. This can be observed by comparing Figure 4.13(a) with 4.10. The data for these figures are obtained from the simulation of the same process and controller, but with different stiction models. The notable features are:

- For a first order plus time delay model, both stiction models show no limit cycle for

the case of pure deadband. Both models show that for limit cycles a certain amount of slip-jump is required.

- If the process contains an integrator in closed loop, both models show limit cycle even in the presence of pure deadband.
- The two models produce identical results for other cases of stiction.

The open loop simulation results for both models look very similar in Figures 4.12 and 4.9. Note that a one-to-one comparison of these figures cannot be made because there is no direct one-to-one relation among the parameters of the empirical data-driven model and that of the physics-based model.

4.5.3 Closed loop behaviour of the model

Closed loop behaviour of the stiction model has been studied for two different cases, namely a concentration loop and a level loop. The concentration loop has slow dynamics with a large dead time. The level loop has only an integrator. The transfer functions, controllers and parameters used in simulation are shown in Table 4.3. Note that the magnitudes of S and J are specified as a percentage (%) of valve input span and output span respectively. The results for each of the loops are discussed in a separate section below.

Table 4.3: Process models, controllers, and data-driven stiction model parameters.

Loop Type	Process	Controller	stiction							
			deadband		undershoot		no offset		overshoot	
			S	J	S	J	S	J	S	J
concentration	$\frac{3e^{-10s}}{10s+1}$	$0.2(\frac{10s+1}{10s})$	5	0	5	2	5	5	5	7
Level	$\frac{1}{s}$	$0.4(\frac{2s+1}{2s})$	3	0	3	3.5	3	3	3	4.5

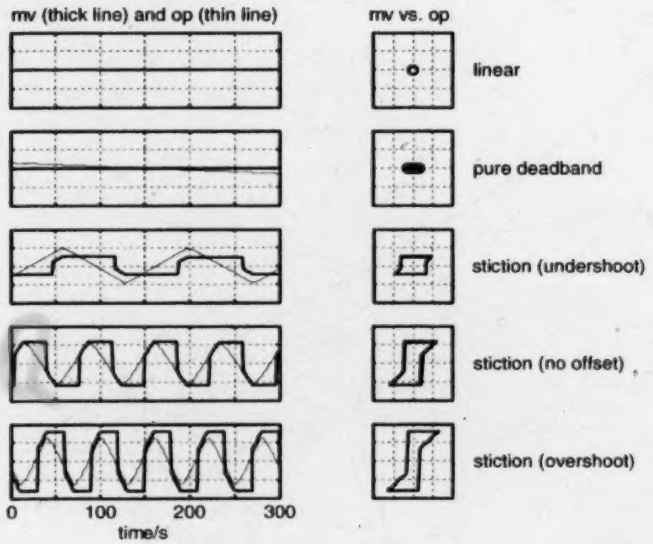
Concentration loop

The transfer function model for this loop was obtained from (Horch and Isaksson, 1998). This transfer function with a PI controller in a feedback closed loop configuration was used for the simulation.

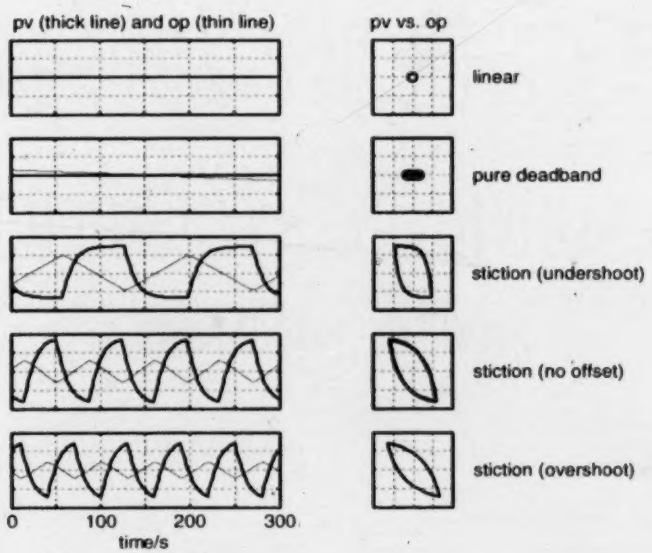
Steady state results of the simulation for different stiction cases are presented in Figures 4.13(a) and 4.13(b). In both figures, thin lines are the controller output. The triangular shape of the time trend of controller output is one of the characteristics of stiction (Horch, 2000). Note that Figure 4.13(a) resembles Figure 4.10 that was generated using the same process and controller in conjunction with the physics-based valve model. In all cases, the presence of stiction causes limit cycling of the process output. In the absence of stiction, there are no limit cycles, which is shown in the first row of Figure 4.13(a). The presence of pure deadband also does not produce a limit cycle. It only adds dead time to the process. This conforms to the findings of (Piipponen, 1996; McMillan, 1995), who stated that the presence of pure deadband only adds dead time to the process and the presence of deadband together with an integrator produces a limit cycle (discussed further in Section 4.6.4). Figure 4.13(a) shows the controller output (*op*) and valve position (*mv*). Mapping of *mv* vs. *op* clearly shows the stiction phenomena in the valve. It is common practice to use a mapping of *pv* vs. *op* for valve diagnosis (see Figure 4.13(b)). However, in this case such a mapping only shows elliptical loops with sharp turn-around points. The reason for the latter is that the *pv-op* map captures not only the nonlinear valve characteristic but also the dynamics of the process, $G(s)$, which in this case is a first order lag plus deadtime. Therefore, if the valve position data are available, one should plot valve position (*mv*) against the controller output (*op*). Except in cases of liquid flow loops where the flow through the valve (*pv*) can be taken to be proportional to valve opening (*mv*), the *pv-op* maps should be used with caution.

A level control loop

The closed loop simulation of the stiction model using only an integrator as the process was performed to investigate the behaviour of a typical level loop in the presence of valve stiction. Results are shown in Figure 4.14(b). The second row of the figure shows that the deadband can produce oscillations. Again, it is observed that if there is an integrator in the process dynamics, then even a pure deadband can produce limit cycles. Otherwise, the cycle decays to zero. The *mv-op* mappings depict various cases of valve stiction. The *pv-op* plots show elliptical loops with sharp turn-around. Therefore, as was noted in an earlier example, the *pv-op* map is not a very reliable diagnostic tool for valve faults in a level loop. A diagnostic technique, developed by Choudhury *et al.* (2004*d*) based on higher order statistical analysis of data, is able to detect and diagnose the presence of stiction in control loops.

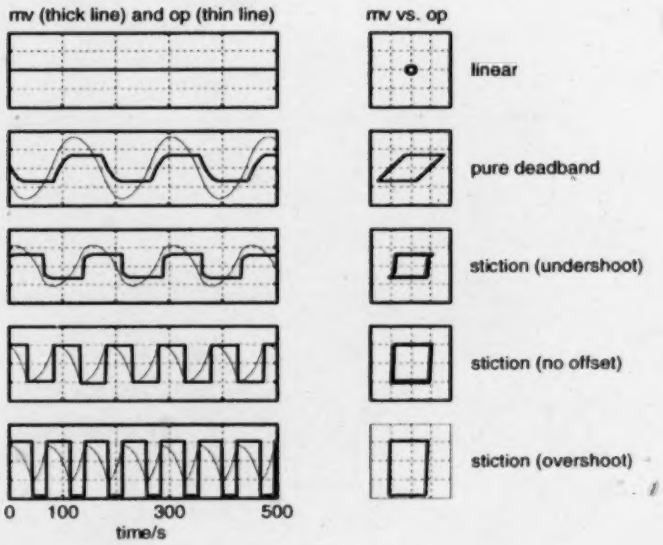


(a) Left: Time trends of *mv* and *op*, Right: *mv-op* mappings

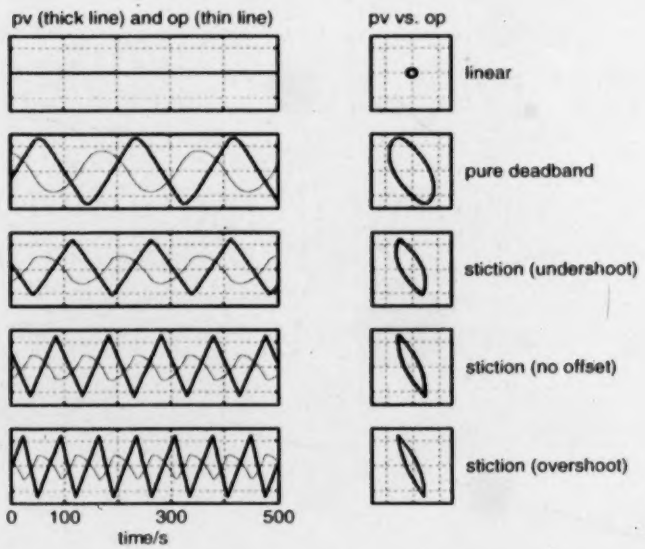


(b) Left: Time trends of *pv* and *op*, Right: *pv-op* mappings

Figure 4.13: *Closed loop simulation results of a concentration loop in presence of the data-driven stiction model.*



(a) Left: Time trends of *mv* and *op*, Right: *mv-op* mappings



(b) Left: Time trends of *pv* and *op*, Right: *pv-op* mappings

Figure 4.14: *Closed loop simulation results of a level loop in presence of the data-driven stiction model.*

4.6 Describing Function Analysis

4.6.1 Introduction

A non-linear actuator with a stiction characteristic may cause limit cycling in a control loop. Further insights into the behaviour of such systems may be achieved through a describing function analysis (Cook, 1986). The non-linearity is modelled by a non-linear gain N. The assumptions inherent in the approximation are that (1) there are periodic signals present in the system and (2) the controlled system is low pass and responds principally to the fundamental Fourier component. The conditions for oscillation in a negative feedback loop arise when the loop gain is -1:

$$G_o(i\omega) = -\frac{1}{N(X_m)} \tag{4.5}$$

where $G_o(i\omega)$ is the open loop frequency response that includes the controlled system and the proportional plus integral controller, and $N(X_m)$ is the describing function that depends on the magnitude of the controller output X_m. When the condition $G_o(i\omega) = -1/N(X_m)$ is met, the system will oscillate spontaneously with a limit cycle. The variation of the quantity $-1/N(X_m)$ with signal amplitude means that signals initially present in the loop as noise can grow until they are large enough to satisfy the equality and hence provide a self-starting oscillation. The solution to the complex equation $G_o(i\omega) = -1/N(X_m)$, if one exists, may be found graphically by superposing plots of $G_o(i\omega)$ and $-1/N$ on the same set of axes.

The aim of describing function analysis is to gain insight into the simulation results and industrial observations presented in this chapter.

4.6.2 An expression for the describing function

The describing function of a non-linearity is:

$$N = \frac{Y_f}{X} \tag{4.6}$$

where X is a harmonic input to the non-linearity of angular frequency ω_o and Y_f is the fundamental Fourier component angular frequency ω_o of the output from the non-linearity. Thus, a Fourier analysis is needed on the output signals shown as bold lines in Figure 4.12. The quantity N depends upon the magnitude of the input X_m. N is complex for stiction non-linearity because the output waveform has a phase lag compared to the input. The

describing function is derived in the Appendix 4.A, where it is shown that:

$$N = -\frac{1}{\pi X_m}(A - iB) \tag{4.7}$$

where

$$A = \frac{X_m}{2}\sin 2\phi - 2X_m \cos\phi - X_m\left(\frac{\pi}{2} + \phi\right) + 2(S-J)\cos\phi, \tag{4.8}$$

$$B = -3\frac{X_m}{2} + \frac{X_m}{2}\cos 2\phi + 2X_m \sin\phi - 2(S-J)\sin\phi \tag{4.9}$$

$$\phi = \sin^{-1}\left(\frac{X_m - S}{X_m}\right) \tag{4.10}$$

4.6.3 Asymptotes of the describing function

Figure 4A.1(a) indicates that there is no output from the non-linearity if $X_m < S/2$. Therefore, the two extreme cases are when $X_m = S/2$ and $X_m \gg S$.

When $X_m \gg S$, then the effects of the deadband and slip-jump are negligible, and the nonlinearity in Figure 4A.1(b) becomes a straight line at 45^o. The output is in phase with the input and $N = 1$. Thus $-1/N(X_m) = -1$ when $X_m \gg S$.

In the limit when $X_m \to S/2$ then the output is as shown in Figure 4.15. The left hand plot shows the output for a slip-jump with no deadband ($S = J$, $d = 0$) while the right hand plot shows a magnified plot of a dead-band with no slip-jump ($S = d$, $J = 0$). In both cases, the output lags the input by one quarter of a cycle. The output is a square wave of magnitude X_m in the $S = J$, $d = 0$ case and the describing function is $N = \frac{4}{\pi}e^{-i\pi/2}$. For the dead-band with no slip-jump ($S = d$, $J = 0$) case, the output magnitude becomes very small. The describing function is $N = \varepsilon e^{-i\pi/2}$, where $\varepsilon \to 0$ as $X_m \to S/2$. Appendix 4.A provides detailed calculations of these results and also shows, for the general case, that the describing function limit when $X_m = S/2$ is:

$$N = \frac{4}{\pi} \times \frac{J}{2}e^{-i\pi/2} \tag{4.11}$$

4.6.4 Insights gained from the describing function

Figure 4.16 shows graphical solutions to the limit cycle equation $G_o(i\omega) = -1/N(X_m)$ for the composition control loop (left panel) and level control loop (right panel) presented earlier. The describing function is parameterized by X_m and the open loop frequency response function of the controller and controlled system is parameterized by ω. Both

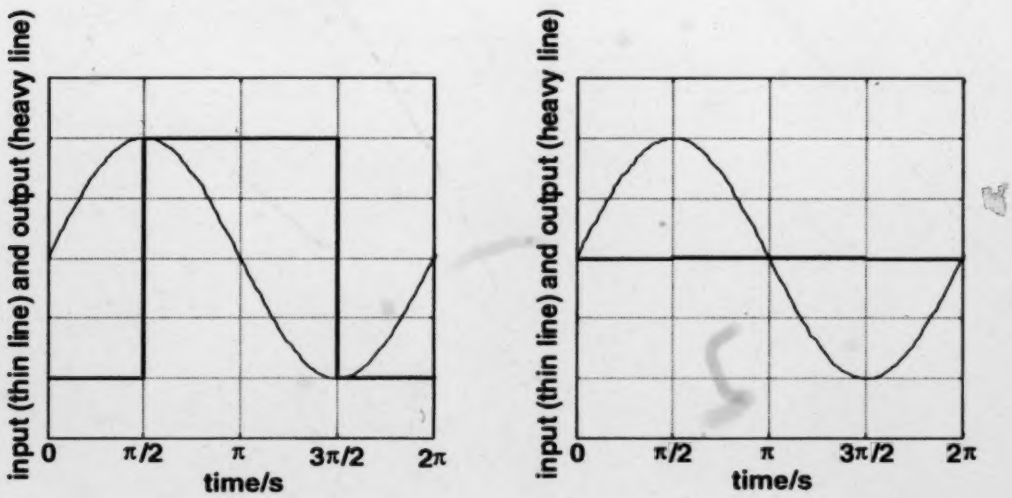


Figure 4.15: *Input (thin line) and output (heavy line) time trends for the limiting case as* $X_m = S/2$. *Left panel: Slip-jump only with* $S = J$. *Right panel: Deadband only with* $J = 0$. *The output in the right plot has been magnified for visualization; its amplitude becomes zero as* X_m *approaches* $S/2$.

systems are closed loop stable and thus intersect the negative real axis between 0 and -1. The plots explain the behaviour observed in simulation.

It is clear from the left plot of Figure 4.16 that there will be a limit cycle for the composition control loop if a slip-jump is present. The slip-jump forces the $-1/N$ curve onto the negative imaginary axis in the $X_m = S/2$ limit. Thus, the frequency response curve of the FOPTD composition loop and its proportional plus integral controller is guaranteed to intersect with the describing function because the integral action means open loop phase is always below $-\pi/2$ (i.e., it is in the third quadrant of the complex plane at low frequency).

Figure 4.16 also shows the $-1/N$ curve for the deadband limit cycle. In the $X_m = S/2$ limit, the curve becomes large, negative and imaginary. The composition loop does not have a limit cycle if the non-linearity is a pure deadband, because the frequency response curve does not intersect the $-1/N$ curve. The lack of a limit cycle in this case has been noted by other authors (Piipponen, 1996; McMillan, 1995).

The level loop with proportional plus integral control has a frequency response for which the phase becomes $-\pi$ at low frequency. The right hand panel of Figure 4.16 shows that it

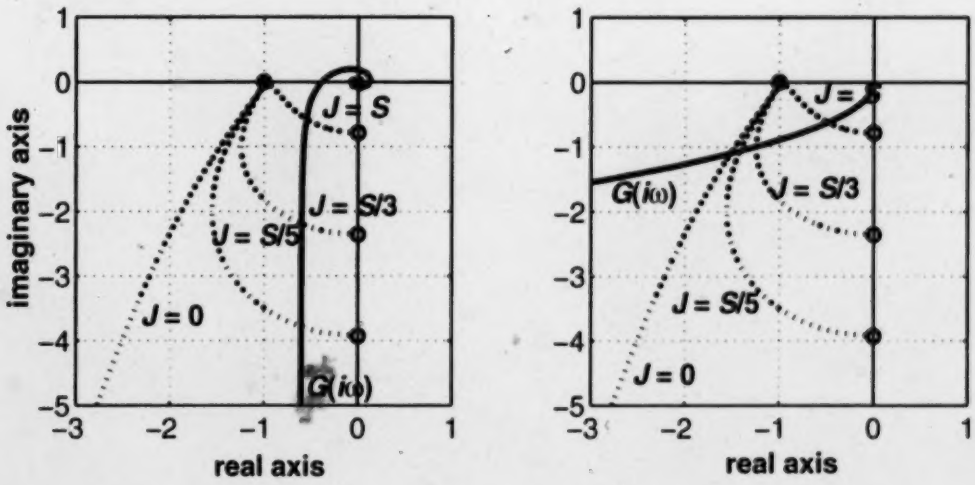


Figure 4.16: *Graphical solutions for limit cycle oscillations. Left panel: Composition control loop. Right panel: level control loop. Dotted lines are the* $-1/N$ *curves and the solid line is the frequency response function.*

will intersect the $-1/N$ curves for the slip-jump cases and also for the pure deadband case. Therefore, a valve with a deadband and no slip-jump can cause a limit cycle oscillation for an integrating process with a P+I controller. The frequency of oscillation is higher and the period of oscillation shorter when the slip-jump is present because the $-1/N$ curves with the slip-jump intersect the frequency response curve at higher frequencies than the $-1/N$ curve for the deadband.

4.7 Conclusion

A generalized definition of valve stiction based on the investigation of real plant data has been proposed. The physics-based model of stiction is difficult to use because of the requirement of knowledge of mass and forces. Therefore, a simple yet powerful data-driven empirical stiction model has been developed. Both closed and open loop results have been presented and validated to show the capability of the model. It is recommended that when using a X-Y plot to analyze valve problems, if *mv* data are available, one should use *mv-op* plot instead of *pv-op*.

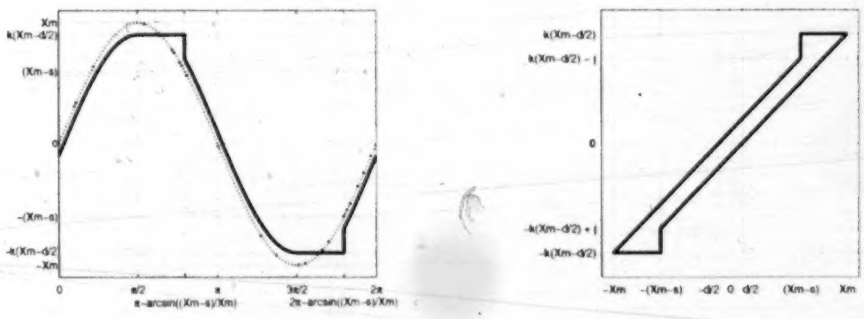
4.A Appendix – Derivation of Describing Function

4.A.1 Expression for the output of the non-linearity

The input-output behaviour of stiction nonlinearity is shown in Figure 4A.1. The output from stiction nonlinearity (i.e., the solid line in the Figure 4A.1(a) is not analytic. It is useful to consider a sine wave input (dotted line in Figure 4A.1(a)) with angular frequency of $1\,\text{rad}\cdot\text{sec}^{-1}$ and period 2π. The output (the solid line in Figure 1(a)) is then:

$$y(t) = \begin{cases} k\left(X_m \sin(t) - \frac{S-J}{2}\right) & 0 \leqslant t \leqslant \frac{\pi}{2} \\ k\left(X_m - \frac{S-J}{2}\right) & \frac{\pi}{2} \leqslant t \leqslant \pi - \phi \\ k\left(X_m \sin(t) + \frac{S-J}{2}\right) & \pi - \phi \leqslant t \leqslant \frac{3\pi}{2} \\ k\left(-X_m + \frac{S-J}{2}\right) & \frac{3\pi}{2} \leqslant t \leqslant 2\pi - \phi \\ k\left(X_m \sin(t) - \frac{S-J}{2}\right) & 2\pi - \phi \leqslant t \leqslant 2\pi \end{cases}$$

where X_m is the amplitude of the input sine wave, S is the deadband plus stickband, J is the slip-jump, $\phi = \sin^{-1}\left(\frac{X_m - S}{X_m}\right)$ and k is the slope of the input-output characteristic in the moving phase (k is assumed to be 1 for a valve).



(a) input-output time trends for stiction

(b) input-output phase plot for stiction

Figure 4A.1: *Describing function analysis for stiction nonlinearity.*

4.A.2 Evaluation of the fundamental Fourier component

The fundamental component of the complex Fourier series is:

$$\frac{1}{2\pi}\int_{t=0}^{2\pi} y(t)\,e^{-it}dt$$

where, after substitution of $\sin(t)=\frac{1}{2i}\left(e^{it}-e^{-it}\right)$:

$$\begin{aligned}\int_{t=0}^{2\pi} y(t)\,e^{-it}dt &= \int_{t=0}^{\pi/2} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)-\frac{S-J}{2}\right)e^{-it}dt + \int_{t=\pi/2}^{\pi-\phi} k\left(X_m-\frac{S-J}{2}\right)e^{-it}dt \\ &+ \int_{t=\pi-\phi}^{3\pi/2} k\left(\tfrac{X_m}{2i}\left(e^{it}-e^{-it}\right)+\tfrac{S-J}{2}\right)e^{-it}dt + \int_{t=3\pi/2}^{2\pi-\phi} k\left(-X_m+\tfrac{S-J}{2}\right)e^{-it}dt \\ &+ \int_{t=2\pi-\phi}^{2\pi} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)-\frac{S-J}{2}\right)e^{-it}dt \qquad (4A.1)\end{aligned}$$

Stating it compactly:

$$\int_{t=0}^{2\pi} y(t)\,e^{-it}dt = T_1+T_2+T_3+T_4+T_5$$

where $T_1=\int_{t=0}^{\pi/2} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)-\frac{S-J}{2}\right)e^{-it}dt$, and so on.

Evaluation term-by-term gives:

$$\begin{aligned}T_1 &= \frac{k}{2}(X_m-S+J)+ik\left(\frac{S-J}{2}-\frac{X_m\pi}{4}\right)\\ T_2 &= -k\left(X_m-\frac{S-J}{2}\right)(1-\sin\phi)-ik\left(X_m-\frac{S-J}{2}\right)\cos\phi\\ T_3 &= k\left(\tfrac{X_m}{4}(1+\cos 2\phi)-\tfrac{S-J}{2}(1+\sin\phi)\right)+ik\left(\tfrac{X_m}{4}\sin 2\phi-\tfrac{X_m}{2}\left(\tfrac{\pi}{2}+\phi\right)+\tfrac{S-J}{2}\cos\phi\right)\\ T_4 &= -k\left(X_m-\frac{S-J}{2}\right)(1-\sin\phi)-ik\left(X_m-\frac{S-J}{2}\right)\cos\phi\\ T_5 &= -k\left(\tfrac{X_m}{4}(1-\cos 2\phi)+\tfrac{S-J}{2}\sin\phi\right)-ik\left(\tfrac{X_m\phi}{2}+\tfrac{d}{2}-\tfrac{X_m}{4}\sin 2\phi-\tfrac{S-J}{2}\cos\phi\right) \qquad (4A.2)\end{aligned}$$

Collecting terms gives the wanted fundamental Fourier component of the output:

$$\frac{1}{2\pi}\int_{t=0}^{2\pi} y(t)\,e^{-it}dt = \frac{1}{2\pi}(B+iA)$$

where $A=\left(k\frac{X_m}{2}\sin 2\phi-2kX_m\cos\phi-kX_m\left(\frac{\pi}{2}+\phi\right)+2k(S-J)\cos\phi\right)$

and $B=-3k\frac{X_m}{2}+k\frac{X_m}{2}\cos 2\phi+2kX_m\sin\phi-2k(S-J)\sin\phi$

The fundamental component of the complex Fourier series of the input sine wave is $X_m/2i$. Therefore, the describing function is:

$$N = \frac{B+iA}{2\pi} \times \frac{2i}{X_m} = -\frac{1}{\pi X_m}(A - iB)$$

4.A.3 Evaluation of limiting cases

There is no output from the non-linearity when $X_m < S/2$. The limiting cases considered are therefore $X_m = S/2$ and $X_m \gg S$.

When $X_m \gg S$, $\phi = \sin^{-1}\left(\frac{X_m - S}{X_m}\right) = \frac{\pi}{2}$, $A = -k\pi X_m$, $B = 0$ and thus $N = k$. This result is to be expected because the influence of the stickband and jump are negligible when the input has a large amplitude and the output approximates a sine wave of magnitude kX_m. The slope of the moving phase for a valve with a deadband is $k = 1$ when the input and output to the non-linearity are expressed as a percentage of full range. Therefore for a valve with stiction, $N = 1$, when $X_m \gg S$.

When $X_m = S/2$ the result depends upon the magnitude of the slip-jump, J. For the case with no deadband ($S = J$), $\phi = -\frac{\pi}{2}$, $A = 0$, $B = -4kX_m$ and $N = -ik\frac{4}{\pi} = k\frac{4}{\pi}e^{-i\pi/2}$. For a valve with $k = 1$, $N = \frac{4}{\pi}e^{-i\pi/2}$. This result describes the situation where the output is a square wave of amplitude X_m lagging the input sine wave by one quarter of a cycle, as shown in Figure 4.15.

For intermediate cases, where both deadband and slip-jump are present such that $|S - J| > 0$, then the $X_m = S/2$ limit gives $\phi = -\frac{\pi}{2}$, $A = 0$, $B = -2kJ$ and $N = -ik\frac{2J}{\pi X_m} = k\frac{2J}{\pi X_m}e^{-i\pi/2}$. For instance, if $J = S/2$ and $k = 1$, then the $X_m = S/2$ limit gives $N = \frac{2}{\pi}e^{-i\pi/2}$ and the output is a square wave of amplitude $X_m/2$ lagging the input sine wave by one quarter of a cycle.

When the non-linearity has a deadband only and no slip-jump ($J = 0$), the describing function has a limit given by $N = \varepsilon e^{-i\pi/2}$ where $\varepsilon \to 0$ as $X_m \to S/2$.

5

Automatic Detection and Quantification of Control Valve Stiction

Stiction is a common problem in spring-diaphragm type valves, which are widely used in the process industry. Although there have been many attempts to understand and detect stiction in control valves, existing methods cannot detect and quantify stiction simultaneously. Conventional invasive methods such as the valve travel test can easily detect stiction, but are expensive and tedious to apply to hundreds of valves. Thus, there is a clear need in the process industry for a non-invasive method that can not only detect but also quantify stiction so that the valves needed repair or maintenance can be identified, isolated and repaired. This chapter describes a method for detecting and quantifying stiction that may be present in control valves using routine operating data obtained from the process. No additional excitation or experimentation of the plant is required. More than a dozen industrial case studies have demonstrated the wide applicability and practicality of this method as a useful diagnostic aid in control-loop performance monitoring.

---

[1] *A brief version of this chapter was published in the proceedings of the DYCOPS 2004 conference (Choudhury et al., 2004c). A regular paper based on this chapter has been submitted to Control Engineering Practice, September 2004. A Patent Application has been filed for the techniques described in this chapter.*

5.1 Introduction

A typical chemical plant has hundreds of control loops. Control performance is important to ensure tight product quality and low cost of the product in such plants. The presence of oscillation in a control loop increases the variability of the process variables thus causing inferior quality products, larger rejection rates, increased energy consumption, reduced average throughput and reduced profitability. The only moving part in a control loop is the control valve. Control valves frequently suffer from problems such as stiction, leaks, tight packing, and hysteresis. Bialkowski (1992) reported that about 30% of the loops are oscillatory due to control valve problems. In a recent study Desborough and Miller (2001) reported that control valve problems account for about one third of the 32% of controllers classified as "poor" or "fair" in an industrial survey (Desborough *et al.*, 2000). If the control valve contains nonlinearities, e.g., stiction, backlash, and deadband, the valve output may be oscillatory, which in turn can cause oscillations in the process output. Among the many types of nonlinearities in control valves, stiction is the most common and one of the long-standing problems in the process industry. It hinders proper movement of the valve stem and consequently affects control loop performance. Stiction can be detected easily using invasive methods such as the valve travel or bump test. However, to apply such invasive methods across an entire plant site is neither feasible nor cost-effective because they are manpower, cost and time intensive in nature.

Although many invasive tests/methods have been suggested (Aubrun *et al.*, 1995; McMillan, 1995; Taha *et al.*, 1996; Wallén, 1997; Sharif and Grosvenor, 1998; Ruel, 2000; Gerry and Ruel, 2001) for analysis and performance of control valves, few non-invasive studies or methods ((Horch, 1999; Rengaswamy *et al.*, 2001; Stenman *et al.*, 2003) have appeared in literature. Horch's method is successful mainly in detecting valve stiction in flow control loops. It cannot be applied to loops involving an integrator or those carrying compressible fluids. The method described by Rengaswamy *et. al.* (2001) depends on the qualitative shape of the time trends of the data, which are often distorted by the presence of noise and disturbance. Also, in the real world, the shape of the time trends of the data is heavily affected by the process and controller dynamics. Stenman *et. al.* (2003) described a model based segmentation method to detect stiction in control valves. This method requires a model of the process and numerous tuning parameters. Obtaining the closed loop model of the process from routine operating data is usually non-trivial. Moreover, all these methods can detect stiction but cannot quantify it. As pointed out by Desborough and Miller (2002), *'a passive or non-invasive method that can reliably and*

automatically classify valve performance in closed loop is desperately needed in process industry'. Clearly, a non-invasive method capable of detecting and quantifying stiction will be useful in the process industry to identify the valves that need maintenance or repair.

An effective non-intrusive data-based monitoring method could reduce the cost of control loop performance maintenance by screening and short-listing those loops and/or valves that need maintenance. This chapter describes a data-based, model-free non-invasive method that can automatically detect and quantify stiction present in control valves. The main contributions of this chapter are:

- A model-free method for detecting and quantifying stiction in control valves from routine operating data is developed. The method does not require the performance of any additional valve travel test or commonly known bump test of the control loop.
- The novel feature of the method is that it can detect and quantify stiction using controlled variable (pv), controller output (op) and set point (sp) data. It does not require valve positioner (mv) data. If mv data are available it is very easy to detect and quantify stiction from the mapping of mv and op. However, this is not the case when only pv, op, and sp data are available because the mapping of pv and op is often confounded by the loop dynamics and disturbances.
- Finally, the algorithm has been fully automated.
- The method is useful in short-listing the valves suffering from stiction from the hundreds or thousands of control valves used in chemical plants or elsewhere. Thus, it contributes to the reduction of plant maintenance costs and increases the overall profitability of the plant.

5.2 Detection of Stiction in Control Valves

In a control loop, a nonlinearity may be present either in the process itself or in the control valve. For our current analysis, we are assuming that the process nonlinearity is negligible in the steady state operating region during which the data has been collected. This is a reasonable assumption because the method works with routine operating data of a control loop under regulatory control. In general, when processes are fairly well regulated at standard operating conditions, the plant can be assumed to behave linearly since a linear controller is capable of satisfactory regulation of the plant.

Control valve nonlinearities include mainly stiction, backlash, deadband and deadzone. Stiction is the most common problem and is the main focus of this study. There are two types of methods for detecting stiction: invasive and non-invasive. Invasive methods require putting the loop in 'manual' and then stroking/travelling the valve over its full travel span. This is now termed the 'valve travel test' in Instrument Society of America (ISA) standards (ISA-75.13-1996; ANSI/ISA-75.05.01-2000). Using this type of test, stiction can be quantified as the amount of change required in the control signal to move the valve from the position where it was stuck. Because it is neither feasible nor cost-effective to test hundreds of valves at a plant site, the non-invasive methods are preferred to invasive methods. Horch's cross-correlation method is popular among the non-invasive methods reported so far for detecting stiction. Horch's method (Horch, 1999; Horch, 2000; Horch *et al.*, 2000) detects stiction with the use of the cross-correlation function between *pv* and *op*. Their method is not applicable to processes containing an integrator, e.g., a level control loop, or for loops carrying compressible media, e.g., steam or air. Their method is useful mainly for flow control loops. Even for flow control loops, it sometimes produces inconclusive results (Desborough and Miller, 2002). Also, if there is a sinusoidal disturbance entering the control loop, the method falsely detects stiction in the control valve (Choudhury *et al.*, 2002). Moreover, none of the existing methods can quantify stiction. Therefore, a new method based on higher order statistics has been developed which can detect as well as quantify stiction and is applicable to all types of control loops. The method first examines the presence of nonlinearity in a control loop. If a nonlinearity is detected, then the process variable (*pv*), set point (*sp*) and controller output (*op*) signals are used to diagnose the possible causes of nonlinearity. The following section describes the method in detail.

5.2.1 Detection of loop nonlinearity

A control loop containing valve nonlinearities often produces non-Gaussian (e.g., a signal with asymmetric distribution) and nonlinear time series, namely process output (*pv*) and controller output (*op*) data. Higher order statistics based nonlinearity assessment can be used as a diagnostic tool for troubleshooting of hardware faults that may be present in the control loop (Choudhury *et al.*, 2002; Choudhury *et al.*, 2004*d*). As described in (Choudhury *et al.*, 2004*d*), the test of Gaussianity and nonlinearity of the control error signal (*sp-pv*) is a useful diagnostic aid for determining the poor performance of a control loop. The test described in Choudhury *et al.* (2003*c*) uses the sensitivity of the normalized

bispectrum or bicoherence to detect the presence of nonlinear interactions in the signal. A distinctive characteristic of a nonlinear time series is the presence of phase coupling such that the phase of one frequency component is determined by the phases of others. Phase coupling leads to higher order spectral features that can be detected in the bicoherence of a signal. The nonlinearity test applied here uses bicoherence to assess the nonlinearity. Bicoherence is defined as:

$$bic^2(f_1,f_2) \triangleq \frac{|B(f_1,f_2)|^2}{E[|X(f_1)X(f_2)|^2]E[|X(f_1+f_2)|^2]} \tag{5.1}$$

where $B(f_1,f_2)$ is the bispectrum at frequencies (f_1,f_2) and is given by

$$B(f_1,f_2) \triangleq E[X(f_1)X(f_2)X^*(f_1+f_2)], \tag{5.2}$$

$X(f_1)$ is the discrete Fourier transform of the time series $x(k)$ at the frequency f_1, $X^*(f_1)$ is the complex conjugate and E is the expectation operator. A key feature of the bispectrum is that it has a non-zero value if there is significant phase coupling in the signal x between frequency components at f_1 and f_2. The bicoherence gives the same information but is normalized as a value between 0 and 1.

Choudhury *et al.* (2004*d*) defined two indices – the Non-Gaussianity Index (*NGI*) and the Non-Linearity Index (*NLI*) – as

$$NGI \triangleq \overline{\hat{bic}^2} - \overline{bic^2}_{crit} \tag{5.3}$$

$$NLI \triangleq |\,\hat{bic}^2{}_{max} - (\overline{\hat{bic}^2} + 2\sigma_{\hat{bic}^2})\,| \tag{5.4}$$

where $\overline{\hat{bic}^2}$ is the average squared bicoherence, $\hat{bic}^2{}_{max}$ is the maximum squared bicoherence, $\sigma_{\hat{bic}^2}$ is the standard deviation of the squared bicoherence and $\overline{bic^2}_{crit}$ is the statistical threshold/critical value obtained from the central chi-square distribution of squared bicoherence. As outlined in (Choudhury *et al.*, 2004*d*), if both *NGI* and *NLI* are greater than zero, then the signal is described as non-Gaussian and nonlinear. The details of the procedure are shown in Figure 5.1. The test can be applied to any time series to check its non-Gaussianity and non-linearity. For a control loop, this test is applied on the error signal (*sp-pv*) to the controller because the error signal is more stationary than *pv* or *op* signal. If the error signal is found to be non-Gaussian and nonlinear, it is inferred that the loop in question exhibits significant non-linearity.

The nonlinearity can be attributed to the control valve under the following assumptions:

- The process is locally linear.

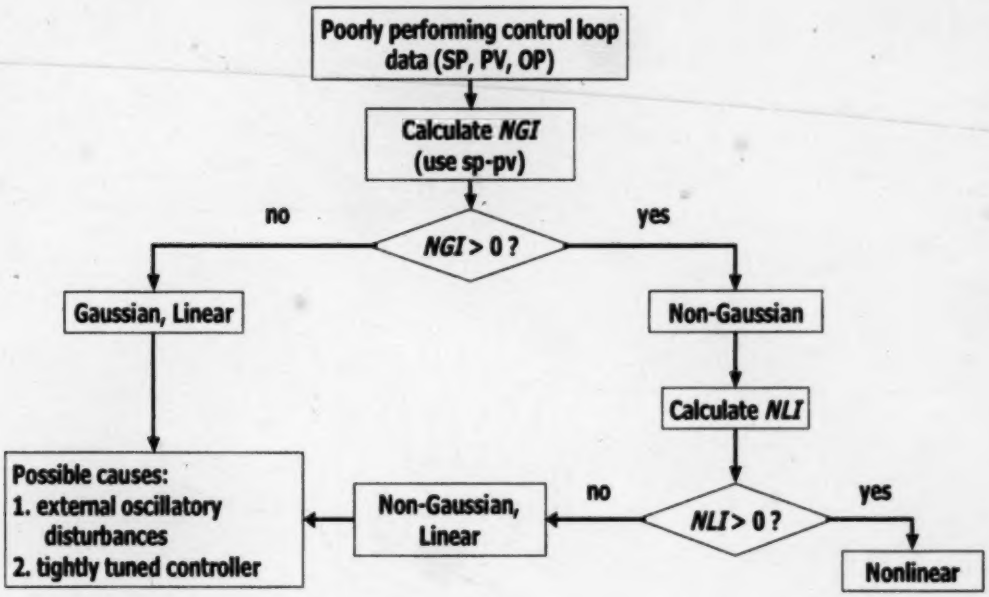


Figure 5.1: *Decision flow diagram of the methodology for the detection and diagnosis of loop nonlinearity.*

- No nonlinear disturbance enters the loop.
- The installed characteristic of the control valve is reasonably linear in the current operating region.

If the disturbance is measurable, then the test can be applied to check the linearity of the disturbance. One may also argue that the valve itself may have a nonlinear characteristic, e.g., a square-root or equal percentage characteristic, which is definitely not a fault. To clarify this situation, a simulation study was performed and the results are already presented in Chapter 3 for equal percentage and square-root valve characteristics in a simple feedback system. It was found that if the movement of the valve stem or the change in input signal to valve is within 15% of the full span (0 to 100%) of the valve travel, then a control loop exhibits linear behavior under steady state regulatory control. It can also be realized by careful investigation of the valve characteristic curves.

Therefore, the higher order statistics based *NGI* and *NLI* indices can easily be calculated

for each loop in an entire plant site and the loops that exhibit nonlinear behavior can be isolated for further diagnosis. After a loop is identified as nonlinear, the causes of nonlinearity should be diagnosed. With the assumptions listed above, it can be concluded that the valve is most likely responsible for the loop nonlinearity. The next problem is to diagnose whether this valve nonlinearity is due to stiction or some other factor. The *pv-op* plot is useful for solving this problem.

5.2.2 Use of *pv-op* plot

The *pv-op* plot has long been used for the detection of valve problems, especially stiction. However, experience shows that this type of method is successful only for a handful number of flow control loops. The use of a *pv-op* plot for detecting valve problems was not successful because it only takes into account the qualitative trend information of the time series, which can be destroyed due to the presence of process dynamics, noise dynamics, disturbances and tightly tuned controllers. In our method, this plot will be used as a second step to diagnose the valve nonlinearity problem, not for the detection of the valve problems. The latter is carried out by using higher statistical based *NGI* and *NLI* indices. If nonlinearity is detected, then the *pv-op* plot is used to diagnose the cause of this nonlinearity. Because of the contamination of real-world data with noise and disturbances, a *pv-op* plot is often unclear and ambiguous, and it is difficult to find any clear information from it. This necessitates the use of a filter to clean the data.

Data Filtering

Since the nonlinearity detection is a frequency domain method, frequency domain based filtering has been chosen here. Upon detection of nonlinearity, the frequencies responsible for significant nonlinear interactions can be determined from the significant peaks in the squared bicoherence plot. Then, a frequency domain Wiener filter is used to obtain those parts of the signal that contribute significantly to signal nonlinearity. Both *pv* and *op* are filtered using a frequency domain Wiener filter. The frequency-domain Wiener filter sets the power in unwanted frequency channels to zero. The filter used here is an approximate realization of a Wiener filter (Press *et al.*, 1986), because a true Wiener filter also requires an estimate of the noise power within the wanted frequency channels, which would then be subtracted from those channels. The detailed design algorithm is given in Thornhill *et al.* (2003), which explains how to deal with aliased frequencies above the Nyquist frequency and constraints on the filter width. The frequency ranges for the filters are selected by

inspecting the peaks in the bicoherence plot. It is preferable to use a large number of data points (e.g., 4096 samples) for the nonlinearity detection algorithm. Filtering is also performed on those data sets. However, the use of such a large number of data points in the pv_f-op_f plot often produces a signature that is difficult to match with any known pattern of valve problems. Therefore, a segment of the data consisting of only several hundred data points should be chosen for the construction of the pv_f-op_f plot. Note pv_f and op_f are the filtered pv and op.

Choosing an appropriate segment of the data

The question that naturally arises is how to select a segment of the data for a useful pv_f-op_f plot. This problem can be resolved by choosing the segment that has regular oscillations because valve problems manifest themselves as limit cycles in the data. Thornhill *et al.* (2003) described a method for the assessment of the period and the regularity of oscillation of a time series. They used the zero-crossings of the auto-covariance function of the time series to estimate the period of oscillation. An oscillation is considered to be regular if the standard deviation of the period of oscillation (σ_{T_p}) is less than one third of the mean value of the period of oscillation ($\bar{T}_p$). The statistic used is:

$$r = \frac{1}{3}\frac{\bar{T}_p}{\sigma_{T_p}} \tag{5.5}$$

A value of r greater than 1 indicates a regular oscillation with a well-defined period. In this work, the filtered controller output signal (op_f) is divided into several segments of user-defined length that can be selected based on the period of oscillation. The segment of op_f corresponding to the highest value of r is used for the pv_f-op_f plot, where pv_f is the corresponding counterpart of op_f. The data segment corresponding to the highest value of r is chosen because valve nonlinearities, e.g., stiction and deadband, are measured as the maximum width of the cycles in the direction of valve input signal in a valve characteristic plot. If the valve positioner data is available (as would be the case for 'smart valves'), a plot of valve output signal (mv) vs. valve input signal (op) can readily be used to quantify stiction. However, in practical cases, valve output or positioner data is seldom available. Therefore, one needs to estimate stiction from the available data of controlled output (pv), controller output (op) and the set point (sp) variables. In this work, stiction is estimated to be the maximum width of the cycles of the pv_f-op_f plot at the direction of op_f. The quantified stiction is termed "apparent stiction" because the actual amount of stiction to be obtained from the mv-op plot may differ from the estimated quantity owing to the effect of

loop dynamics on the controlled variable, *pv*, in particular, the effect of the controller to compensate or fight stiction.

5.3 Quantifying Stiction

The detection and diagnosis algorithm can identify stiction in a large number of control valves. Some valves may have an acceptable level of stiction, while others may have severe stiction that may demand immediate maintenance. Therefore, it is important to be able to quantify stiction so that a list of sticky valves can be prepared in order of their maintenance priority. It is well known that the presence of stiction in control valve in a control loop produces limit cycles in the controlled variable (*pv*) and the controller output (*op*) (Hagglund, 1995; Horch, 1999; Ruel, 2000; Rengaswamy *et al.*, 2001). For such a case, if *pv* is plotted against *op*, cyclic patterns are found in the resulting *pv*-*op* plot. A large number of such plots can be found in (Choudhury *et al.*, 2004*a*; Choudhury *et al.*, 2004*b*), where stiction models were used in a closed loop SISO system to produce data for these plots. An ellipse can be fitted to such a pv_f-op_f plot. The pv_f-op_f plot together with either of the following two methods can be used to quantify stiction in unit of the *op* signal. Note that there is no need to scale the data.

5.3.1 Clustering technique

Clustering is a method for dividing scattered groups of data into several groups. Because the *pv*-*op* plot for a control loop with a sticky valve exhibits elliptic patterns, the data corresponding to a narrow strip along the mean of *pv* and parallel to the *op* axis can be collected (see Figure 5.2(c)) and then used for quantifying stiction with the help of c-means or fuzzy c-means clustering techniques. A detailed description of these methods is represented in Appendix - 5.A. The amount of stiction can be estimated from the absolute value of the difference between the x co-ordinates of the centers of the two clusters. If the final centers of the clusters are (op_1, pv_1) and (op_2, pv_2), then the amount of stiction is determined using the following expression:

$$Apparent\ \ Stiction = |op_1 - op_2| \tag{5.6}$$

5.3.2 Using a fitted ellipse

An ellipse in the least square sense can be fitted in the pv_f-op_f plot and then used to quantify stiction. For details about the theory of ellipse fitting, refer to Appendix - 5.B.

Because apparent stiction is defined as the maximum width of the ellipse in the *op* direction, the distance between two points lying at the intersections of the ellipse and a line parallel to the *op* axis and passing through the center of the ellipse will be the amount of stiction present in the loop. For any point $P(x,t_2)$ in the $X-Y$ co-ordinate system (see Figure 5.2(f)), Equation 5B.16 can be solved using Equations 5B.18 and 5B.19. This gives the X co-ordinate of points A and P in Figure 5.2(f):

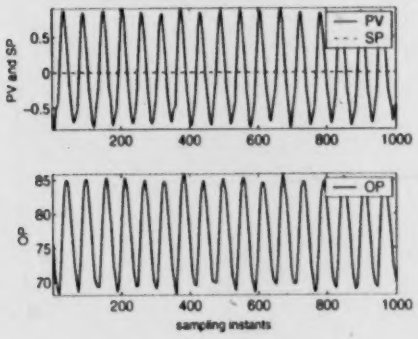
$$x = t_1 \pm \frac{mn}{\sqrt{(m^2\, sin^2\theta + n^2\, cos^2\theta)}} \tag{5.7}$$

where (t_1,t_2) is the center of the fitted ellipse, m and n are the length of the major and minor axes of the fitted ellipse respectively, and θ is the angle of rotation of the ellipse. Therefore, the amount of stiction (length of AP in Figure 5.2(f)) can be obtained using the following expression

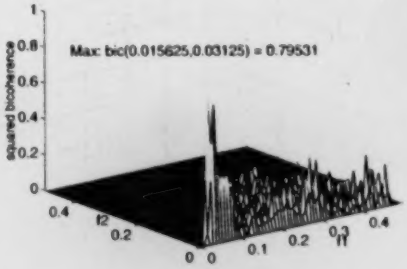
$$Apparent\ \ Stiction\ =\ AP\ =\ \Delta x\ = \frac{2mn}{\sqrt{(m^2\, sin^2\theta + n^2\, cos^2\theta)}} \tag{5.8}$$

5.4 An Illustrative Example

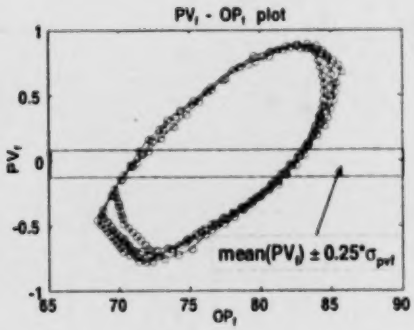
The objective of this section is to explain the sequence of steps in the proposed method with a detailed presentation of an industrial example. This example represents a level control loop in a power plant, which controls the level in a condenser located at the outlet of a turbine by manipulating the flow rate of the liquid condensate from the condenser. In total, 8640 samples for each tag were collected at a sampling rate of 5 seconds. Figure 5.2(a) shows the time trends for level (*pv*), set point (*sp*) and the controller output (*op*). The loop shows oscillatory behavior. For the bicoherence calculation, 4096 data points were used. Figure 5.2(b) shows the squared bicoherence plot corresponding to the controller error signal (*sp-pv*). The values of *NGI* and *NLI* were found to be 0.04 and 0.61 respectively, indicating the presence of significant loop nonlinearity. From the bicoherence plot 5.2(b), it can be seen that frequencies in the range 0.001 to 0.1 are the most significant frequencies of the signal responsible for nonlinear interactions. Therefore, the *pv* and *op* signals were filtered using a Wiener filter with frequency boundaries at 0.001 and 0.1. Using the method of Thornhill *et al.* (2003) described earlier, it was found that the controller output signal



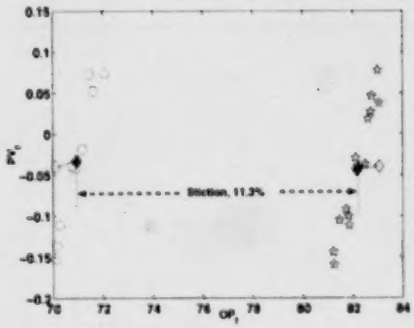
(a) Time trends



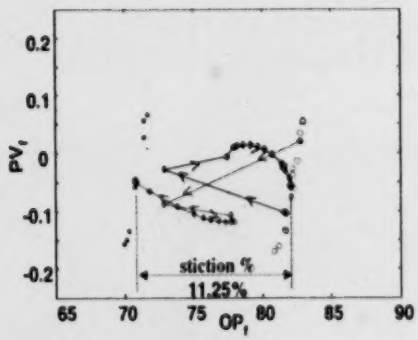
(b) bicoherence plot



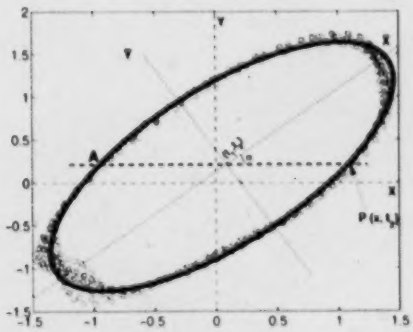
(c) *pv-op* plot



(d) c-means clustering



(e) fuzzy c-means clustering



(f) fitted ellipse

Figure 5.2: *Data analysis results for an industrial level control loop. The biocherence plot shows large peaks indicating nonlinearities in the loop. The subplots (c), (d) and (e) demonstrate c-means and fuzzy c-means method of quantifying stiction. The ellipse fitting technique of quantifying stiction is shown in the subplot (f). Approximately, 11% stiction*

was showing regular oscillations with an average period of 19.78 sampling instants and the maximum r value 10.5 for a segment length of 200 data points. The maximum r value corresponds to the 2801 to 3000 samples.

The pv_f-op_f plot is found to be useful for isolating nonlinearity. Thus, the filtered pv_f and op_f corresponding to this segment is plotted in Figure 5.2(c), which shows excellent elliptical patterns, indicating valve stiction. Figure 5.2(d) demonstrates the c-means clustering technique used in quantification of the stiction. The points denoted by empty and filled diamonds are the initial and final centers of the clusters respectively. This method quantifies the amount of stiction in this loop as 11.3%. In contrast, Figure 5.2(e) illustrates the use of fuzzy c-means clustering in the quantification of stiction. The trajectories followed by the centers of the clusters during the iteration stages are shown by lines with diamonds directed with arrows. The final centers are again in solid diamonds. The amount of stiction estimated by this method is 11.25%. Figure 5.2(f) shows the algebraic ellipse fitting technique, and the amount of stiction estimated using this method is 11.40%. All three methods produced identical results, with practically tolerable limits of deviation from each other.

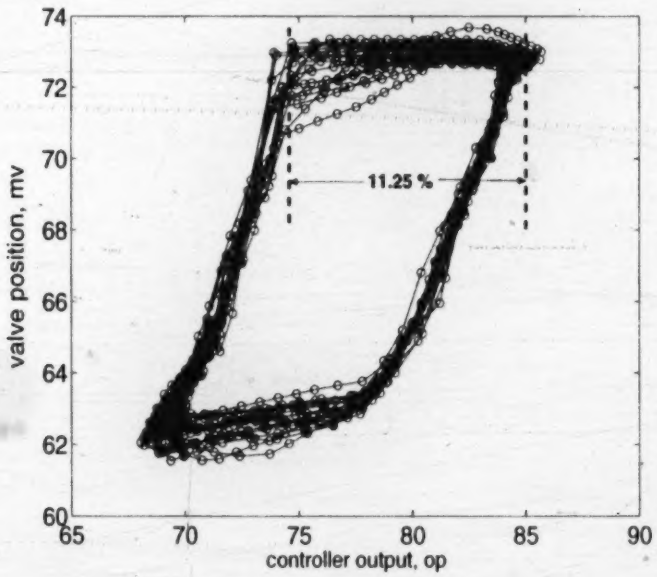


Figure 5.3: *Valve Position (mv) versus Controller Output (op) plot. This plot confirms that the amount of stiction was correctly estimated in Figure 5.2.*

Validation of the Results of the Illustrative Example: After the results of the analysis were sent to plant engineers, they confirmed that this loop was suffering from stiction. For this loop, the valve positioner data were made available. Figure 5.3 shows the actual valve position (mv) vs. controller output (op) plot. This plot clearly shows that the valve was sticking during the change of its direction. From this plot, the amount of stiction can be estimated as 11.25%, which is in agreement with the results obtained from the proposed methods.

5.5 Automatic Detection and Quantification of Stiction

In order to apply the proposed method to a large number of industrial control loops, it must be automated. The following is the description of the automation steps.

1. Calculate *NGI* and *NLI* for the control error signal (sp-pv). If both indices are greater than 0, then go to the following step. Otherwise, STOP. Nonlinearity is not a problem. The poor performance may be caused by tight tuning, detuned controller or external oscillatory disturbances (refer to Figure 5.1).

2. After nonlinearity is detected, obtain the frequency (f_1, f_2) corresponding to the maximum bicoherence peak in step 1. Note that all frequencies are normalized such that the sampling frequency is 1. Suppose f_1=min(f_1 , f_2) and f_2=max(f_1 , f_2).

3. The boundaries of a Winer filter can be obtained from [ω_L = max(0.001 , f_1-0.05), ω_H = min(0.5, f_2+0.05)]. Note that 0.05 is subtracted or added from the frequencies in order to ensure that the exact location of the significant peak does not fall on the filter boundaries.

4. Filter pv and op data to obtain pv_f and op_f.

5. Obtaining the segment of the data with regular oscillations

   (a) Choose a segment length L, say $L = 1000$.

   (b) Divide the op_f data into segments of length L. Here op_f is chosen instead of pv_f because often the op signal is less noisy than the pv signal.

   (c) Calculate r and T_p for each segment of op_f data.

   (d) Obtain $r_{max} = max\ (r)$.

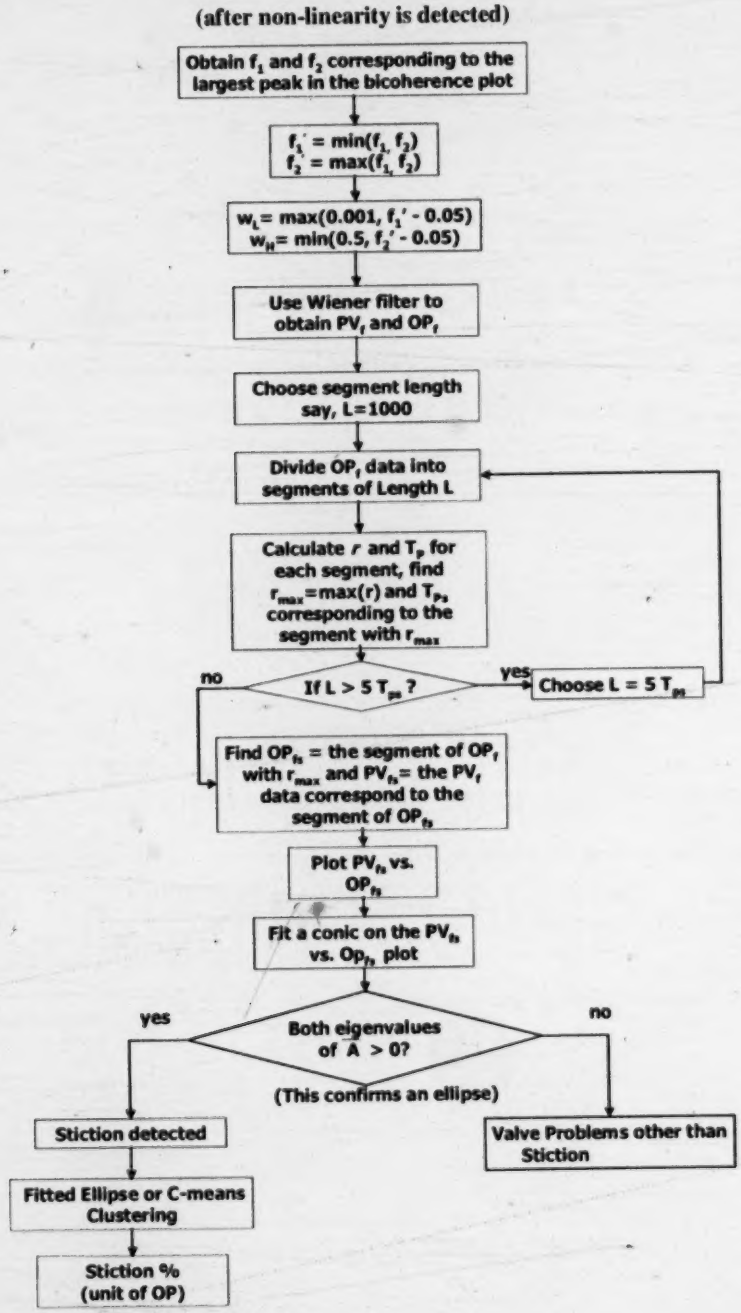

(after non-linearity is detected)
Obtain f_1 and f_2 corresponding to the largest peak in the bicoherence plot
$f_1' = \min(f_1, f_2)$
$f_2' = \max(f_1, f_2)$
$w_L = \max(0.001, f_1' - 0.05)$
$w_H = \min(0.5, f_2' + 0.05)$
Use Wiener filter to obtain PV_f and OP_f
Choose segment length say, L=1000
Divide OP_f data into segments of Length L
Calculate r and T_p for each segment, find $r_{max}=\max(r)$ and T_{ps} corresponding to the segment with r_{max}
If $L > 5\,T_{ps}$?
yes
Choose $L = 5\,T_{ps}$
no
Find OP_{fs} = the segment of OP_f with r_{max} and PV_{fs}= the PV_f data correspond to the segment of OP_{fs}
Plot PV_{fs} vs. OP_{fs}
Fit a conic on the PV_{fs} vs. Op_{fs} plot
Both eigenvalues of A > 0?
yes
no
(This confirms an ellipse)
Stiction detected
Valve Problems other than Stiction
Fitted Ellipse or C-means Clustering
Stiction %
(unit of OP)

Figure 5.4: *Decision flow diagram of the methodology for the detection and quantification of valve stiction.*

(e) Obtain T_{ps}, which is equal to the T_p of the segment of op with r_{max}.

(f) If $L > 5\,T_{ps}$, then choose $L = 5\,T_{ps}$ and go to step (b).

(g) Now, op_{fs} is the segment of the op_f data that corresponds to the r_{max} and pv_{fs} is the portion of the pv_f data that corresponds to op_{fs}.

6. Use the pv_{fs} and op_{fs} data to get pv_f vs. op_f plot.

7. Fit a conic to the selected pv_{fs} and op_{fs} data. If both eigenvalues of the $\bar{A}$ (see Appendix -5.B) matrix are greater than zero, then the pv_f vs. op_f plot is an ellipse. Otherwise, print a message, "Not an ellipse — Other valve problems, not stiction".

8. Quantify stiction using Formula 5.8 or 5.6, depending on the method chosen for stiction quantification.

The above mentioned automatic detection and quantification of stiction algorithm has been summarized in Figures 5.1 and 5.4. Figure 5.1 shows the automatic detection of loop nonlinearity. After the nonlinearity is detected, Figure 5.4 can be used to quantify stiction.

5.6 Simulation Results

5.6.1 Diagnosis of stiction

This section demonstrates the applicability of the proposed method for the detection and quantification of valve stiction through a known simulated case of stiction. A simple single-input, single-output (SISO) system in a feedback control configuration (Figure 5.5) was used for generating simulated data. The first order process with time delay is given by the following transfer function:

$$G(z^{-1}) = \frac{z^{-3}(1.45 - z^{-1})}{1 - 0.8z^{-1}} \tag{5.9}$$

The process is operating under regulatory control with a PI controller. A random walk disturbance generated by integrating random noise was added to the process. The signal to noise ratio defined by the ratio of the variance of the controlled output (pv) to the variance of the random noise was 6. The simulation was performed for 6000 sampling intervals. To remove the effect of transients, the first several hundred data points were discarded and the last 4096 points of the error signal to the controller (sp-pv) were analyzed to detect nonlinearity present in the system.

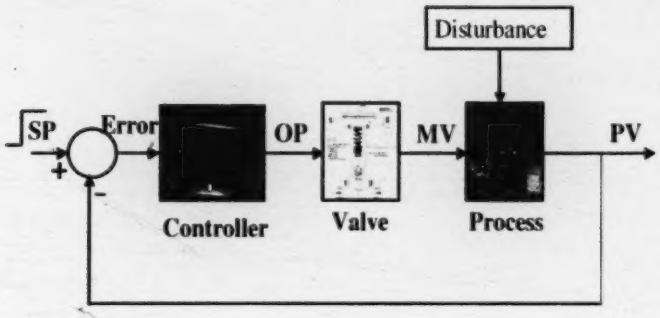
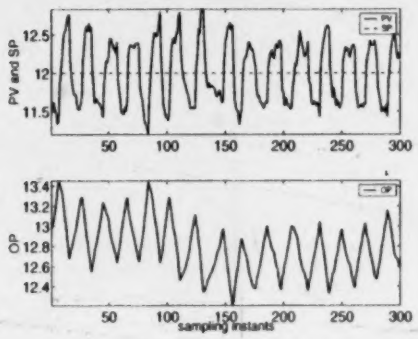
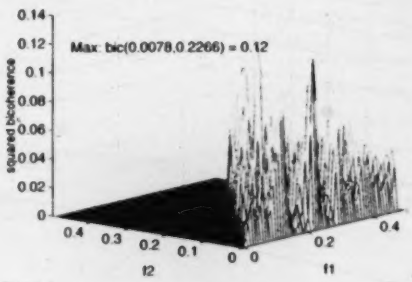


Figure 5.5: *Block diagram of a simple SISO process with stiction nonlinearity in the valve.*

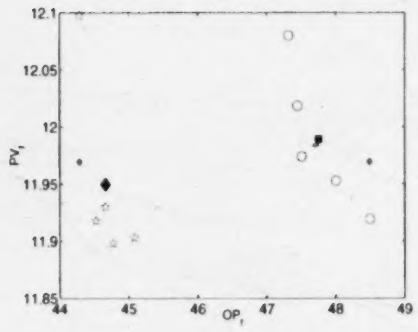
A stiction model developed by Choudhury *et al.* (2004a, 2004b) was used to introduce the stiction behavior of the control valve in the closed loop process. A 3% stiction ('S') with 1% slip-jump ('J') were used in simulation. Note that in order to initiate limit cycles or oscillations in a simple first order time delay process in the presence of valve stiction, a set point change at the beginning of the simulation is required. The process is then left to operate under regulatory control. Figure 5.6(a) shows the time trends of *pv*, *op*, and *sp*. The presence of stiction produces oscillations in the process. The values of *NGI* and *NLI* are 0.01 and 0.06, clearly detecting the presence of nonlinearity in the process signal. The bicoherence plot (Figure 5.6(b)) shows that the frequency range of interest from a nonlinearity point of view is [0.001 0.28] (using the steps of automation section). After performing Wiener filtering, the segments of the pv_f and op_f data corresponding to the segment of op_f that has the highest oscillation index were chosen to obtain the pv_f-op_f plot. C-means clustering, fuzzy c-means clustering and fitted ellipse technique were used to quantify stiction. The amount of stiction obtained from these techniques were estimated to be 2.98%, 2.98% and 3.2% respectively, which is in agreement with the amount of stiction used in simulation. Thus, all these methods are capable of correctly quantifying the amount of stiction present in a control loop. However, for the sake of brevity, only c-means clustering and fitted ellipse technique will be used in subsequent sections. The reader can choose any of these methods to quantify apparent stiction.



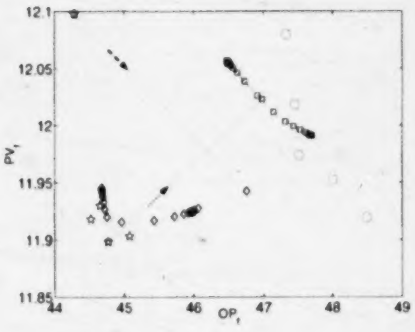
(a) Time trends



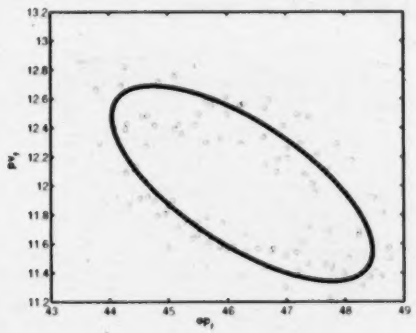
(b) bicoherence plot



(c) c-means clustering



(d) fuzzy c-means clustering

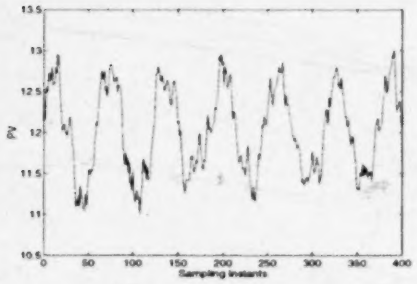


(e) fitted ellipse

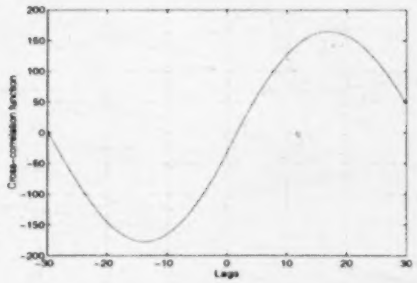
Figure 5.6: *Results of detection and quantification of stiction in a simulated data set.*

5.6.2 Diagnosis of a sinusoidal disturbance

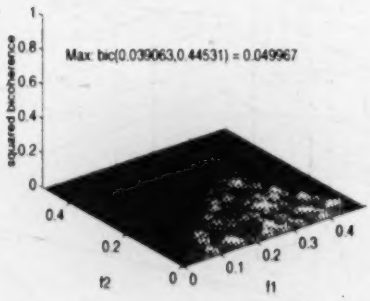
Oftentimes an unmeasured oscillatory disturbance (for example, a sine wave) can initiate cycles in the controlled and manipulated variables. This can be misdiagnosed as a valve problem. This example illustrates the effect of addition of a sinusoidal disturbance with amplitude 2 and frequency 0.01 Hz to the process. Note some measurement noise were also introduced (refer to Figure 5.5).



(a) Time trends



(b) Cross-correlation plot



(c) bicoherence plot

Figure 5.7: *Stiction detection results for a simulated data set when an external sinusoidal oscillatory disturbance plus noise are entering the loop. The cross-correlation method detects stiction, even though there was no stiction (plot (b)). The bicoherence based method correctly shows that there is no nonlinearity this loop (subplot(c)).*

The stiction model was absent in this simulation study. Thus, the diagnosis results should not detect stiction or any other nonlinearity. The time trend of the controlled variable (*pv*) in Figure 5.7(a) shows the oscillatory behavior of the process output. Horch's cross-correlation test (Horch, 1999) shows an odd correlation function, indicating possible valve stiction (see Figure 5.7(b)). However, the higher order statistical test developed provides a measure for $NGI \approx 0$, indicating a linear loop. The bicoherence plot for the error signal to the controller is shown in Figure 5.7(c). The flatness of the bicoherence plot confirms the linearity of the loop.

5.6.3 Diagnosis of root cause of a propagated disturbance

This case demonstrates root cause diagnosis of propagation of oscillation(s) from one loop to another. This simulation case has been formulated by feeding output of a concentration loop as a disturbance to a level control loop. The approximate transfer function and controllers for both loops are given below.

For concentration Loop:

$$G(s) = \frac{3e^{-10s}}{10s+1} \qquad C(s) = 0.2(1+\frac{1}{10s}) \tag{5.10}$$

For level Loop:

$$G(s) = \frac{e^{-2s}}{4s} \qquad C(s) = 0.1(1+\frac{1}{10s}) \tag{5.11}$$

The simulation block diagram for this study is shown in Figure 5.8. The data driven stiction model was used in the composition loop with $S = 3$ and $J = 2$ and the valve in the level control loop was free from stiction. In reality, it may describe a scenario where the outlet of a mixing chamber in a composition loop is used to feed another processing unit (e.g., a stock tank) and the level of the unit is controlled by a PI controller.

Results of the analysis of these two loops are shown in Figure 5.9. As shown in figure, both loops are oscillating. The *NGI* and *NLI* values for the composition loop are 0.004 and 0.57 respectively, and for the level loop, they are $NGI = 0.002$ and $NLI = 0.49$. The amount of apparent stiction detected in the composition loop is 0.70% whereas in the level loop it is 1%. The method falsely detects stiction in the level loop. The method developed here assumes that the disturbance(s) entering the loop isarelinear. For the level loop, this assumption has been violated. For application in industry, where oscillations propagate plant-wide, it is recommended that the stiction diagnosis method should be used in conjunction with information from the process flow sheet. For example, for this case

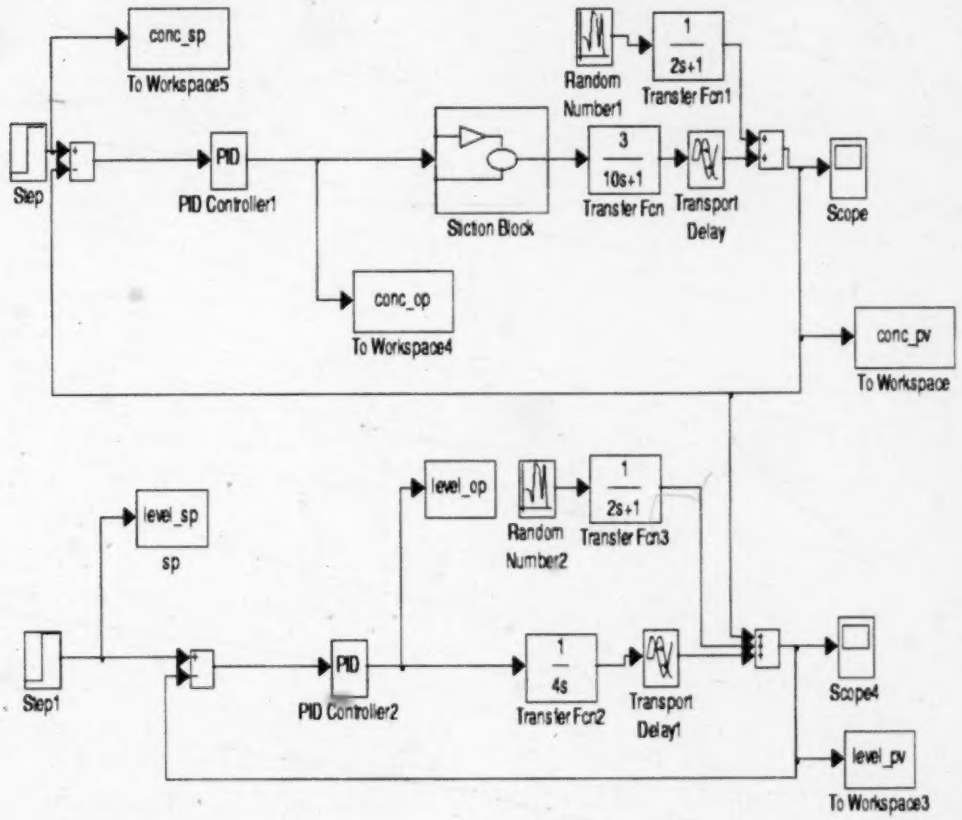
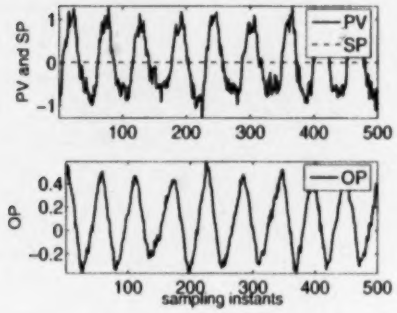
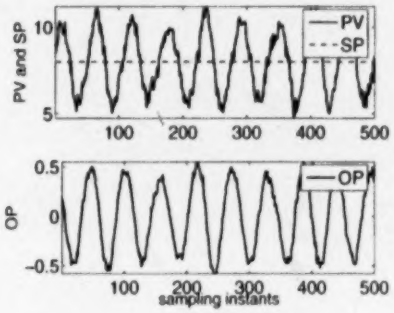


Figure 5.8: *Simulation Block diagram for stiction induced oscillation propagation study.*

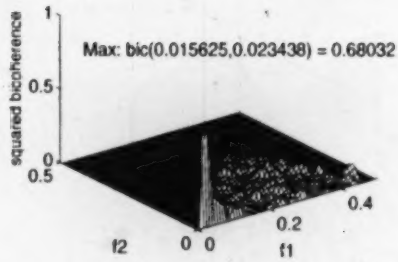
these two loops are connected and stiction is detected in both loops. Since there is no recycle, it is most likely that the source of nonlinearity due to stiction is located in the upstream loop, i.e., in the composition loop. The *NGI* and *NLI* indices are also higher for the composition loop, confirming this as the root cause of the nonlinearity. Other stiction detection algorithms (Horch, 1999; Rengaswamy *et al.*, 2001; Stenman *et al.*, 2003) would also produce false positive results for this kind of propagated disturbance (s) because they also do not consider the interaction among the loops.



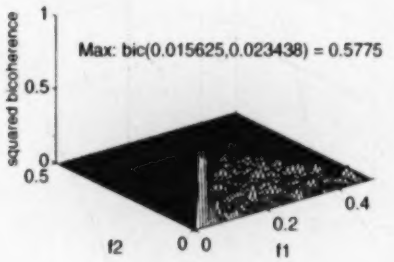
(a) Time trends



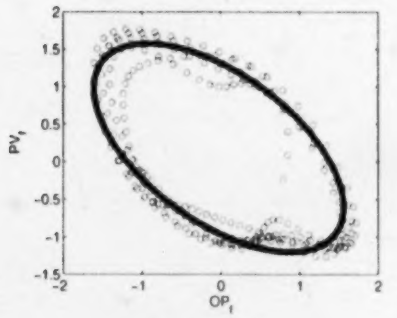
(b) Time trends



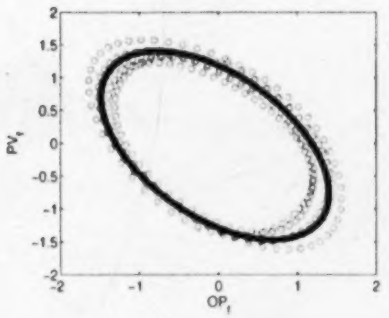
(c) bicoherence plot



(d) bicoherence plot



(e) fitted ellipse



(f) fitted ellipse

Figure 5.9: *Analysis results for root cause diagnosis of a propagated disturbance. Result for the composition loops with a sticky valve (left). Results for the level loop affected by the disturbance from the composition loop (right).*

5.7 Industrial Case Studies

The objective of this section is to evaluate the proposed method on a number of selected control loop obtained from different types of process industries. For each loop, set point (*sp*), controlled output (*pv*) and controller output (*op*) data were available. Unless otherwise stated, a data length of 4096 was used for the squared bicoherence calculation for each case. The time trends of these variables, the squared bicoherence plot, the c-means clustering plot, and the fitted ellipse plot for each loop are presented. The numerical results for all loops are provided in Table 5.1. These data were analyzed without prior knowledge of the control valve problems, and the results of the analysis were confirmed later by the plant personnel.

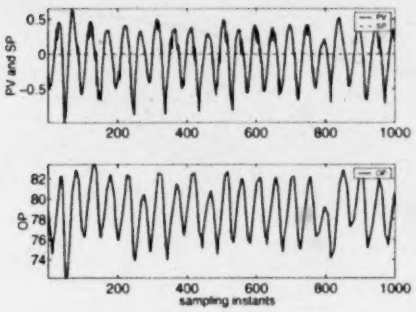
Table 5.1: Numerical Results for the Industrial Loops Analyses

Loop No.	Loop Type	NGI	NLI	ω_L	ω_H	T_p	r	Apparent Stiction %	
								c-means	ellipse
1	Level	0.10	0.40	0.001	0.08	50	22	4.2	4.3
2	Level	-0.02	—	—	—	95	3.5	—	—
3	Flow	0.01	0.55	0.001	0.08	45	8.4	0.35	0.33
4	Temp	0.003	0.19	0.004	0.28	125	6.5	1.00	1.14
5	Pressure	0.02	0.17	0.01	0.25	12.2	12.2	11.00	11
6	Composition	0.02	0.38	0.01	0.15	28.3	11.6	1	1
7	Flow	0.006	0.38	0.004	0.14	59	4.6	—	—

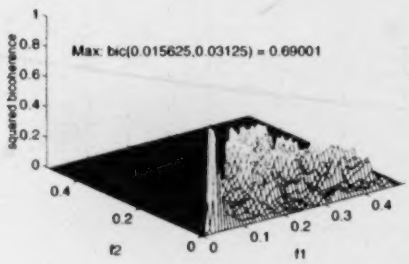
5.7.1 Loop 1 : A level loop

This is a level control loop in the same power plant described in the illustrative example. It also controls the level of condensers located at the outlet of a turbine by manipulating the flow rate of the liquid condensate. Figure 5.10(a) shows the time trend of the *sp*, *pv*, and *op* data. Figure 5.10(b) shows the squared bicoherence plot. The values of NGI and NLI were 0.1 and 0.40. These clearly indicate that nonlinearity is a problem for this loop. From the position of the maximum peak at the bicoherence plot, the frequency range for the Wiener filter was obtained following the steps described in the automation section (Section 5.5). The frequency band for the filter is [0.001 0.08] (1000 samples/cycle to

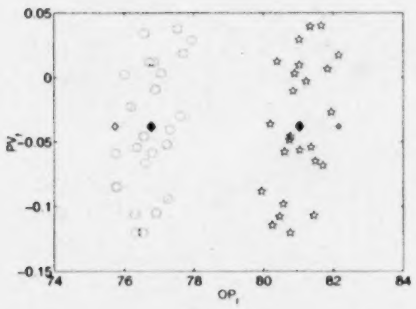
12.5 samples/cycle). The average period of oscillation(s) was 50 samples for the controller output signal. The segment of the data corresponding to the maximum oscillation index (the magnitude of r was 22 for this case) was selected to quantify stiction. Both c-means clustering and fitted ellipse techniques estimate the amount of apparent stiction as 4%. The c-means clustering plot is shown in Figure 5.10(c), while Figure 5.10(d) shows the fitted ellipse technique for quantifying stiction.



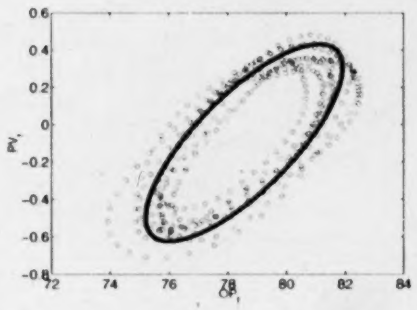
(a) Time trends



(b) bicoherence plot



(c) c-means clustering, filled diamonds are the final centers of the clusters

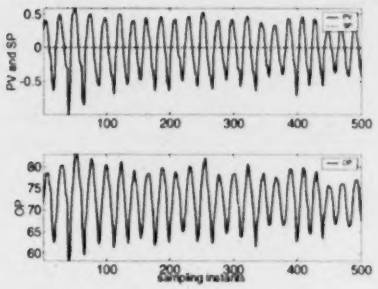


(d) fitted ellipse is shown as heavy curve

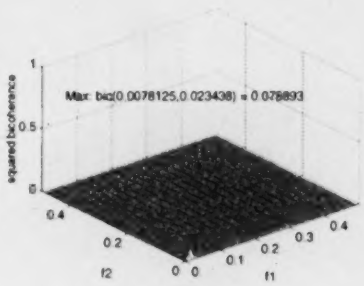
Figure 5.10: *Stiction detection and quantification results for an industrial level control loop data. Significantly large peaks in the bicoherence plot detects nonlinearity in the loop. The presence of ellipse confirms stiction. The width of the ellipse or c-means clustering method quantifies stiction.*

5.7.2 Loop 2: A linear level control loop

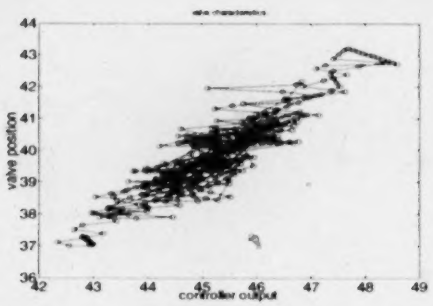
This loop is another level control loop in the same power plant described in the illustrative example. It also controls the level of a condenser located at the outlet of a different turbine by manipulating the flow rate of the liquid condensate. Figure 5.11(a) shows the time trend of the *sp*, *pv*, and *op* data. Figure 5.11(b) shows the squared bicoherence plot for the control error signal. The magnitude of NGI was -0.02, clearly indicating nonlinearity is not a problem for this loop. Figure 5.11(c) shows the valve positioner (*mv*) vs. controller output (*op*) plot. From this figure, it is obvious that the valve shows a linear response.



(a) Time trends



(b) bicoherence plot

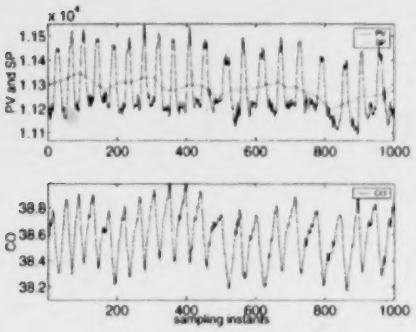


(c) *mv*-*op* plot

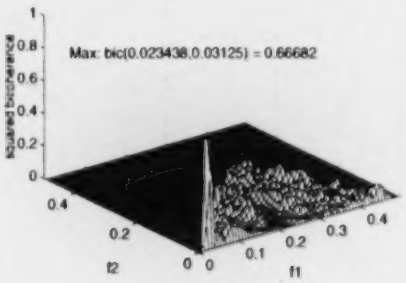
Figure 5.11: *Results of the analysis of another industrial level control loop data. The flatness of the bicoherence plot shows the linearity of this level control valve, which was confirmed in the valve position versus valve input plot (subplot (c)).*

5.7.3 Loop 3: A flow control loop

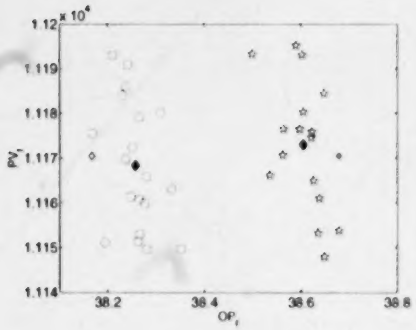
This is a flow control loop obtained from a refinery. The results of the analysis of this loop are shown in Figure 5.12 and also in the third row of the Table 5.1. The presence of a small amount of stiction (0.35% for loop 3) was causing a large amplitude oscillation (see the magnitude of *pv* axis in Figure 5.12(a)) in this loop.



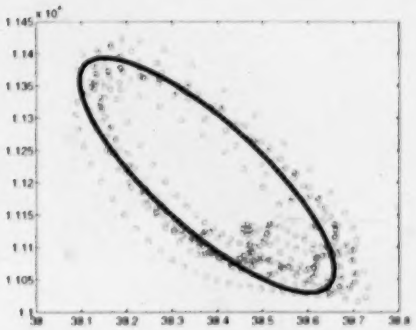
(a) Time trends



(b) bicoherence plot



(c) c-means clustering, filled diamonds are the final centers of the clusters

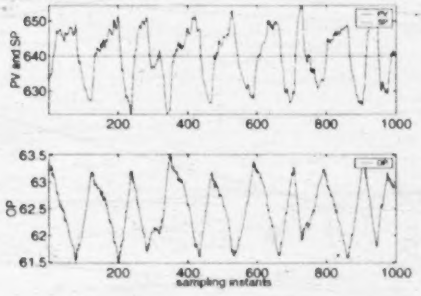


(d) fitted ellipse is shown as heavy curve

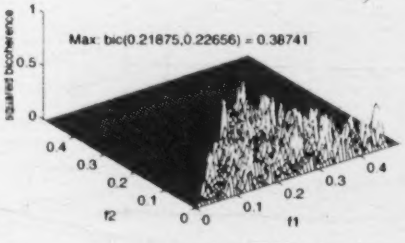
Figure 5.12: *Results of the analysis of a refinery flow control loop data. Significant peaks in the bicoherence plot show the nonlinearities in the loop. The presence of ellipse in pv-op plot confirms stiction.*

5.7.4 Loop 4: A temperature control loop

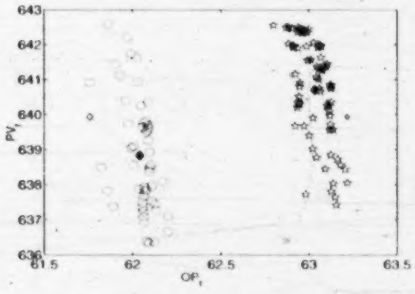
This is a temperature control loop on a furnace feed dryer system at the Tech-Cominco mine plant in Trail, British Columbia, Canada. The temperature of the dryer combustion chamber is controlled by manipulating the flow rate of natural gas to the combustion chamber. The top left plot of Figure 5.13 shows time trends of temperature (*pv*), set point (*sp*) and controller output (*op*). It shows clear oscillations both in the controlled variable (*pv*) and the controller output. The other results are presented in the sixth row of the Table 5.1 and in Figure 5.13. The amount of stiction found in this loop was approximately 1%.



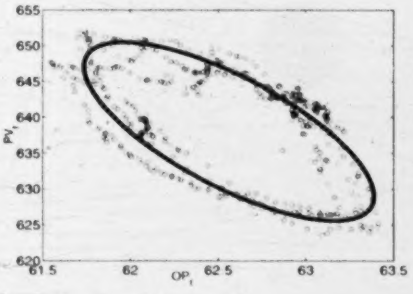
(a) Time trends



(b) bicoherence plot



(c) c-means clustering, filled diamonds are the final centers of the clusters

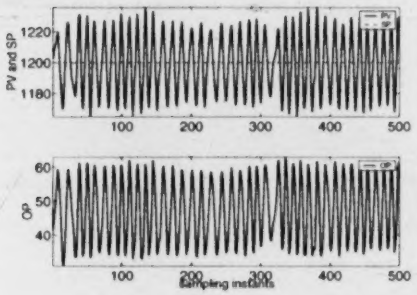


(d) fitted ellipse is shown as heavy curve

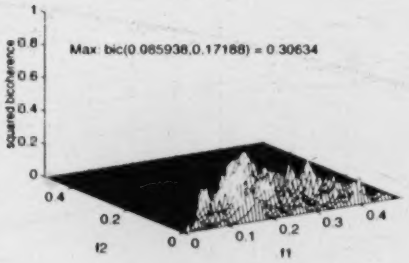
Figure 5.13: *Results of stiction detection and quantification in an industrial furnace dryer temperature control loop. Approximately 1.1% stiction was detected in the natural gas flow control valve.*

5.7.5 Loop 5: A pressure control loop

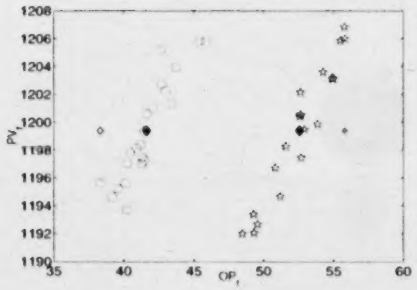
This is a pressure control loop in a refinery plant. This data set had only 1500 data points collected at 20 second sampling intervals. The time trends in Figure 5.14 show oscillations with 12.2 samples in both *pv* and *op* variables. The detailed results of the analysis are presented in Figure 5.14 and in the 5^{th} row of Table 5.1. The apparent stiction present in the valve was approximately 11%.



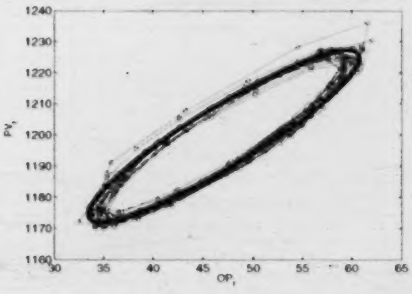
(a) Time trends



(b) bicoherence plot



(c) c-means clustering, filled diamonds are the final centers of the clusters

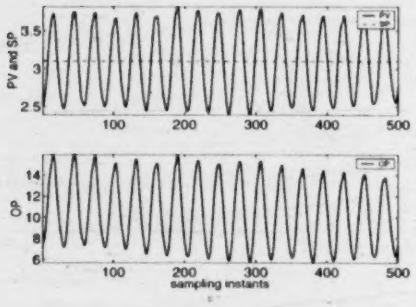


(d) fitted ellipse is shown as heavy curve

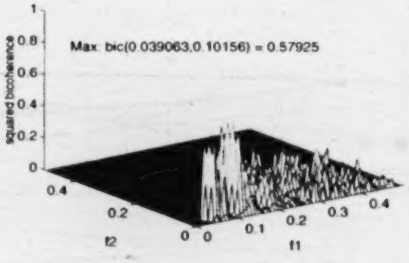
Figure 5.14: *Results of the analysis of a refinery pressure control loop data. Approximately 11% stiction was present in the control valve.*

5.7.6 Loop 6: A composition control loop

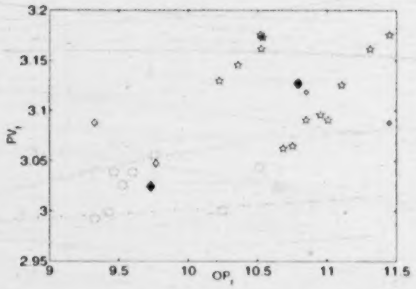
This describes a concentration control loop. The data set contains 1100 points collected at 1 second sampling intervals. The time trends in Figure 5.15(a) show oscillations with 28.3 samples in both *pv* and *op* variables. The detailed results of the analysis are presented in Figure 5.15 and also in the 6^{th} row of Table 5.1. The apparent stiction present in the valve of this concentration control loop was approximately 1%.



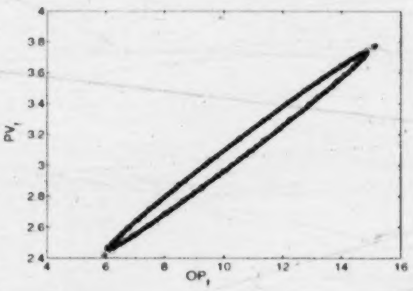
(a) Time trends



(b) bicoherence plot



(c) c-means clustering, filled diamonds are the final centers of the clusters

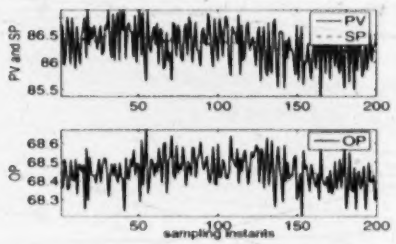


(d) fitted ellipse is shown as heavy curve

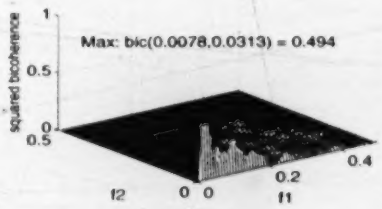
Figure 5.15: *Results of the analysis of an industrial composition control loop data. Approximately 1% stiction was present in the control valve.*

5.7.7 Loop 7: A cascaded flow control loop

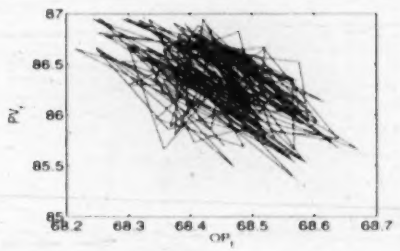
This loop represents a flow control loop cascaded with a level control loop of a drum in an ethylene plant. At a sampling rate of 1 sample/min, 7200 data points were collected. Time trends for the *pv* and *op* variables for the flow control loop are shown in Figure 5.16(a). The detailed results of the analysis are presented in Figure 5.16 and also in the 7^{th} row of Table 5.1. The bicoherence plot shows presence of nonlinearity in this loop. The absence of an elliptical pattern in the *pv-op* plot (Figure 5.16(c)) indicates that this nonlinearity is not due to valve stiction. The probable source of this nonlinearity is likely to be nonlinear disturbances entering this loop or valve problems other than stiction or nonlinearities in the process itself.



(a) Time trends



(b) bicoherence plot



(c) *pv-op* plot

Figure 5.16: *Results of the analysis of an industrial cascaded flow loop data. The bicoherence plot shows the nonlinearity present in this loop. The absence of an ellipse in the pv-op plot (subplot (c)) indicates that this nonlinearity is not due to valve stiction.*

5.8 Online Compensation for Stiction

After a sticky valve is identified the best solution is to repair it, which may require taking the valve out-of-service. In most cases, it is not possible to take the sticky valve out-of-service for its immediate maintenance because of the absence of a bypass line. A method that can help continuing operation of the plant until next planned shutdown with minimum effects of stiction can be useful for the process industry. There are several online methods to deal with stiction (Bergstrom and Dumont, 2003; Armstrong-Hélouvry *et al.*, 1994; Hatipoglu and Ozguner, 1998; Kayihan and Doyle III, 2000; Hagglund, 2002; Tao *et al.*, 2002). For online compensation of stiction Gerry and Ruel (2001) suggested the use of a PI controller where the integral action has variable strength. Hagglund (2002) suggested a method of adding short pulses called 'knockers' to the control signal to compensate stiction in pneumatic valves. As claimed by Hagglund (2002), this 'knocker' friction compensation method has been patented and implemented in industrial controllers. In (Bergstrom and Dumont, 2003), a special adaptive controller in combination with a PID controller has been suggested for online stiction compensation. It requires a special function for switching between the adaptive controller and the PID controller. They did not apply this to industrial settings. Other methods to compensate stiction in control valves can be found in (Armstrong-Hélouvry *et al.*, 1994; Hatipoglu and Ozguner, 1998; Kayihan and Doyle III, 2000; Tao *et al.*, 2002; Amin *et al.*, 1997).

5.9 Conclusions

A non-invasive method for detecting and quantifying stiction in control valves has been presented in this chapter. The method first detects nonlinearity in a control loop by the use of the sensitivity of the normalized bispectrum or bicoherence to the nonlinear interactions that may be present in the control error signal. If nonlinearity is detected, then pv and op signals are filtered using frequency domain Wiener filter to obtain filtered pv_f and op_f signals. If an ellipse can be fitted satisfactorily onto the pv_f-op_f plot, this is a signature of valve stiction. Then C-means clustering, fuzzy c-means clustering or fitted ellipse techniques can be used to automatically quantify the amount of stiction. The method has been extensively evaluated on simulated as well as industrial data sets.

5.A Appendix Clustering Techniques

5.A.1 C-means clustering

In this method, data are partitioned into C number of initial clusters. Then proceeding through all data points, each point is assigned to the nearest cluster (in terms of Euclidean distance). The centroids for the cluster receiving the new item and for the cluster losing the item is re-calculated. This procedure is repeated until no more reassignments take place. For details, refer to (Johnson and Wichern, 1998). This method requires the initialization of the centers of the clusters. In our case, there are only two clusters, and the centers can be specified as [min(op_f), mean(pv_f)] and [max(op_f), mean(pv_f)] calculated from the data obtained along the stripe in the pv_f-op_f plot (see Figure 5.2(c)).

5.A.2 Fuzzy C-means clustering

The fuzzy c-means clustering method as described in (Dulyakarn and Rangsanseri, 2001; Bezdek, 1981) works based on the minimization of the following objective function.

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m \| X_j - V_i \|^2, \quad 1 \le m < \infty \tag{5A.1}$$

where U is a fuzzy c-partition of the data set, V is a set of K prototypes, m is any real number greater than or equal to 1, u_{ij} is the degree of membership of X_j in the cluster i, X_j is the j^{th} observation of the d-dimensional measured data, V_i is the d-dimension center of the cluster, and $\| * \|$ is any norm expressing the similarity between a measured data and the center. Through an iterative optimization of Equation 5A.1, fuzzy partition is carried out with the update of membership of u_{ij} and the cluster centers V_i using the following equations.

$$u_{ij} = \frac{1}{\sum_{k=i}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \tag{5A.2}$$

$$V_i = \frac{\sum_{j=1}^{n} u_{ij}^m X_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{5A.3}$$

where, d_ij is the Euclidean distance between the observation $X(j)$ and the center of the cluster V_i. The criteria to end the iteration is $\max(|u_{ij} - \hat{u}_{ij}|) < \varepsilon$, where ε is a number

close to zero (e.g., 10^{-5}) and $\hat{u}_{ij}$ is the membership number at the previous iteration step. In this study, the number of clusters is two and 10^{-5} was used as the magnitude of ε.

5.B Appendix Fitting an ellipse

Assume that given data points are **op** and **pv**, where

$$\mathbf{op} = [op(1), op(2), \cdots, op(N)]^T \tag{5B.1}$$

$$\mathbf{pv} = [pv(1), pv(2), \cdots, pv(N)]^T \tag{5B.2}$$

$$\tag{5B.3}$$

Starting with the general equation of conics, the equation of ellipse is developed. The equation for any conic in the ordinary $X-Y$ co-ordinate is given by the following equation

$$a_1 {x_1}^2 + a_{12} x_1 x_2 + a_2 {x_2}^2 + b_1 x_1 + b_2 x_2 + c = 0 \tag{5B.4}$$

or,

$$\Phi\Theta = 0 \tag{5B.5}$$

where $\Phi = [{x_1}^2 \;\; x_1 x_2 \;\; {x_2}^2 \;\; x_1 \;\; x_2 \;\; 1]$, $\Theta = [a_1 \;\; a_{12} \;\; a_2 \;\; b_1 \;\; b_2 \;\; c]^T$, x_1 corresponds to data from the *op* signal and x_2 corresponds to data from the *pv* signal. Now, for a given data set, the above equation can be solved as constrained least squares problem: $||\Phi\Theta|| = min$ subject to $||\Theta|| = 1$.

Often real-world data sets require a linearly shifted and rotated conic. Therefore, there is a need to fit a rotated and shifted conic in a transformed co-ordinate $\bar{X} - \bar{Y}$ (see Figure 5.2(f). Equation 5B.4 can be rewritten as:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c = 0 \tag{5B.6}$$

with $\mathbf{x} = [x_1, x_2]^T$, $\mathbf{A} = [a_1 \;\; a_{12}/2; \;\; a_{12}/2 \;\; a_2]$, $\mathbf{b} = [b_1 \;\; b_2]^T$. Note that $\mathbf{A}$ is symmetric and positive definite. Let us use the following equation for the transformation of the equation in the new co-ordinate system $\bar{X} - \bar{Y}$,

$$\mathbf{x} = \mathbf{Q}\bar{\mathbf{x}} + \mathbf{t} \tag{5B.7}$$

where $\mathbf{Q}$ is the matrix for rotational transformation and $\mathbf{t}$ is the vector in the original $X-Y$ co-ordinate for a linear shift of the conic. Using Equation 5B.7, the equation of the conic in the transformed co-ordinate can be written as:

$$\bar{\mathbf{x}}^T \mathbf{Q}^T \mathbf{A} \mathbf{Q} \bar{\mathbf{x}} + (2\mathbf{t}^T \mathbf{A} + \mathbf{b}^T) \mathbf{Q} \bar{\mathbf{x}} + \mathbf{t}^T \mathbf{A} \mathbf{t} + \mathbf{b}^T \mathbf{t} + c = 0 \tag{5B.8}$$

This can be rewritten in the following simplified form

$$\bar{\mathbf{x}}^T\bar{\mathbf{A}}\bar{\mathbf{x}}+\bar{\mathbf{b}}^T\bar{\mathbf{x}}+\bar{c}=0 \tag{5B.9}$$

where

$$\bar{\mathbf{A}} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} \tag{5B.10}$$

$$\bar{\mathbf{b}}^T = (2\mathbf{t}^T\mathbf{A}+\mathbf{b}^T)\mathbf{Q} \tag{5B.11}$$

$$\bar{c} = \mathbf{t}^T\mathbf{A}\mathbf{t}+\mathbf{b}^T\mathbf{t}+c \tag{5B.12}$$

Now, $\mathbf{Q}$ can be chosen in a way so that $\bar{\mathbf{A}} = diag(\lambda_1, \lambda_2)$. One approach is to choose $\mathbf{Q}$ as the eigen-vector matrix obtained from the eigenvalue decomposition of the matrix $\mathbf{A}$. If the conic is an ellipse with its center at the origin of the new co-ordinate $\bar{X}-\bar{Y}$, then in Equation 5B.9,

$$\bar{\mathbf{b}} = \mathbf{0} \tag{5B.13}$$

Therefore, Equation 5B.9 can be simplified as

$$\lambda_1\bar{x_1}^2+\lambda_2\bar{x_2}^2+\bar{c}=0 \tag{5B.14}$$

or,

$$\frac{\bar{x_1}^2}{\sqrt{\frac{-\bar{c}}{\lambda_1}}}+\frac{\bar{x_2}^2}{\sqrt{\frac{-\bar{c}}{\lambda_2}}}=1 \tag{5B.15}$$

or,

$$\frac{\bar{x_1}^2}{m^2}+\frac{\bar{x_2}^2}{n^2}=1 \tag{5B.16}$$

where,

$$m=\sqrt{\frac{-\bar{c}}{\lambda_1}}, \qquad n=\sqrt{\frac{-\bar{c}}{\lambda_2}} \tag{5B.17}$$

The lengths of the axes of the ellipse will be invariant to the transformation. Therefore, in the original co-ordinate $X-Y$, the lengths of the axes of the ellipse are $2m$ and $2n$, respectively. The center of the ellipse is at $\mathbf{t}$, which can be calculated from $\mathbf{t} = -0.5\mathbf{A}^{-1}\mathbf{b}$ (obtained using Equation 5B.13). The angle of rotation of the ellipse (θ, measured anti-clockwise from the positive X axis) can be calculated using any of the eigenvectors. Since the eigenvectors are of unit length, Equation 5B.7 can be written as

$$x_1 = \bar{x_1}\,cos\theta-\bar{x_2}\,sin\theta+t_1 \tag{5B.18}$$

$$x_2 = \bar{x_1}\,sin\theta+\bar{x_2}\,cos\theta+t_2 \tag{5B.19}$$

6

Impact of Data Compression and Quantization on Data-Driven Process Analyses

Modern chemical plants are extensively instrumented and automated. One manifestation of increasing automation and accessibility is easy availability of process information on the desktop. Information access is typically available as a result of extensive data logging and process archiving. Historical data are an invaluable source of information. In chemical industrial practice, data are often compressed using various techniques, e.g., box car, backward slope, swinging door, PLOT, wavelet, etc. before storing them in a historian. Compression degrades data quality and induces nonlinearity. This chapter focuses on the problems of data quality degradation and nonlinearity induction due to compression, and automatic detection and quantification of the compression present in the archived data. Finally, the problem of quantization in the process data is discussed and an automatic procedure to detect and quantify quantization is presented.

---

[1] *A full paper based on this chapter has been published as Thornhill, N. F., Choudhury, M. A. A. S. and Shah, S. L., "The Impact of Compression of Data-Driven Process Analysis", Journal of Process Control, 14, June 2004, pp. 389-398. This chapter is a significantly expanded version of the published paper with more examples, discussions on the compression algorithms and quantization problems.*

6.1 Introduction

The motivations for data compression include reduction of the costs of storage of historical data and reduction of cost of transmission of process data through a telecommunications link. For instance, in pharmaceutical manufacturing the regulatory authorities demand long term storage of manufacturing records, in which case the cost of storage media would be a consideration. In off-shore oil production, the cost is in the satellite linkage to an on-shore headquarter. The trend towards remote monitoring of their installed systems by technology vendors also requires data transmission through a telecommunications link.

However, data compression has hidden costs if the data become unsuitable for their intended purposes. The operation of restoring the original signal from the archived data is called reconstruction. Once the data have been compressed they lose information and the reconstructed trends are deficient in various ways compared to the originals. End uses of the reconstructed data may be very different. They include:

- Calculation of daily statistics such as daily means, daily standard deviations;
- Averaging for data reconciliation and mass balancing;
- Archiving data trends for subsequent high fidelity reconstruction;
- Data smoothing by removal of high frequency noise;
- Feature extraction and recovery of events.

For example, data transmitted from an off-shore production platform are used to determine daily totals of oil flow into the pipeline for taxation purposes, while remote monitoring of a model predictive controller at a refinery may need high fidelity data for identification of a dynamic process model.

The objectives of this study are to provide new insights into the impact of data compression on data-driven plant performance analysis and to recommend how much compression can be tolerated. The findings that compression causes problems may seem obvious in retrospect, but there appears to have been no systematic study reported in the literature to date. An automated means of detecting the severity of compression is also presented. Application of the algorithm during the data pre-processing phase of a plant audit means less time is wasted in evaluation of unsuitable data. It also avoids the loss of credibility of the methods and their practitioners that might arise if the wrong conclusions were to be drawn from bad data.

6.2 Motivating example

Figure 6.1 shows a data set from a historian typical of those from which engineers and consultants wish to extract useful information (courtesy of Celanese Canada Inc.). The straight line segments characteristic of industrial data compression can be seen in many of the time trends. It will be shown in Section 6.6 that compression factors of up to 94 were in use. The original uncompressed data were lost forever when they were compressed and archived, and it is now impossible to determine what features were lost. Later sections will show that most of these data trends are too compressed and that data-driven process analysis would, if attempted, give a misleading indication of the results that the original data would have provided.

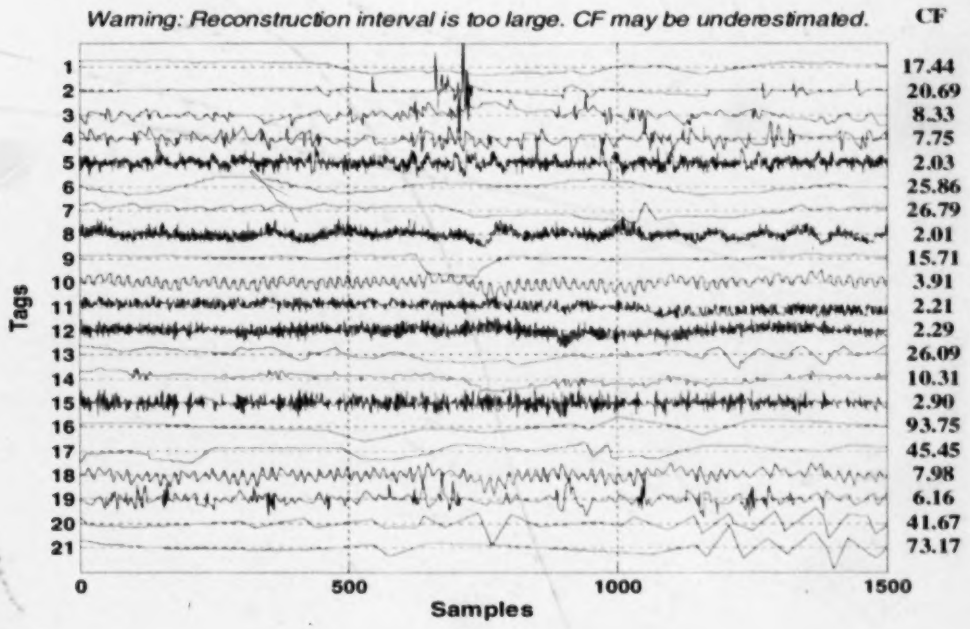


Figure 6.1: *An industrial data set with compression in some tags. Time trends are mean centered and normalized. CF as shown on the right column is an estimate of the compression factor (as defined in Section 6.9.2).*

6.3 Methods

Compression using piecewise linear trending is widely used in industrial data historians. For instance, AspenTech described an adaptive method based upon the box-car/backward slope method (Aspentech, 2001), while OSI state that their PI data historian uses a swinging door compression algorithm (OSI Software Inc., 2002).

6.3.1 Overview of data compression

There is extensive literature on methods of compressing images, speech and text (Watson, 1993). Compression techniques for ECG (electrocardiogram) signals are also at an advanced stage (Crowe *et al.*, 1992; Karczewicz and Gabbouj, 1997). The motives in ECG are like those for process data compression in regard to transmission of the ECG signals by telephone. Some developments in data compression have arisen from that field, for example wavelet compression has moved from ECG to process applications.

Compression techniques can be divided into two main functional groups: direct methods and transform methods.

6.3.2 Direct methods

The direct methods (also known as piecewise linear trending methods) are used in industrial practice because these can be applied in real-time to spot data. Mah *et al.* (1995) and Watson *et al.* (1998) have provided comparative reviews of various compression methods. Mah *et al.* (1995) compared piecewise linear trending methods and introduced a new method (PLOT). Watson *et al.* (1998) studied piecewise linear trending as well as wavelet and Fourier compression. They also introduced a method using vector quantization for process applications and discussed its benefits.

A direct method makes the archiving decision in real time, concurrently with the capture of data from the process. Various types of algorithms are used for data compression in process industries. Examples include box car (BC), backward slope (BS), combined box car and backward slope (BCBS), and Swinging door (sdoor). These algorithms (Hale and Sellars, 1981; Bristol, 1990) use heuristic rules to decide whether to archive a spot value. Those rules are tuned to achieve the capture of exceptions and linear trends. They reconstruct a data trend as a series of linear segments connecting the archived spot values of the data. A brief description of these algorithms are given below.

Box Car (BC) Algorithm

In this algorithm, the current data point is compared to the past recorded data point. If the two values differ by more than or equal to the recording limit or the compression deviation set for this variable, the point immediate prior to the current data point is archived. The details of the algorithm are shown in Figure 6.2. This method is helpful only for steady processes. If there is a ramp type change (e.g., the level of a tank as it is being filled, or the process shifts frequently from one operating point to another), then this method does not provide a high level of data compression.

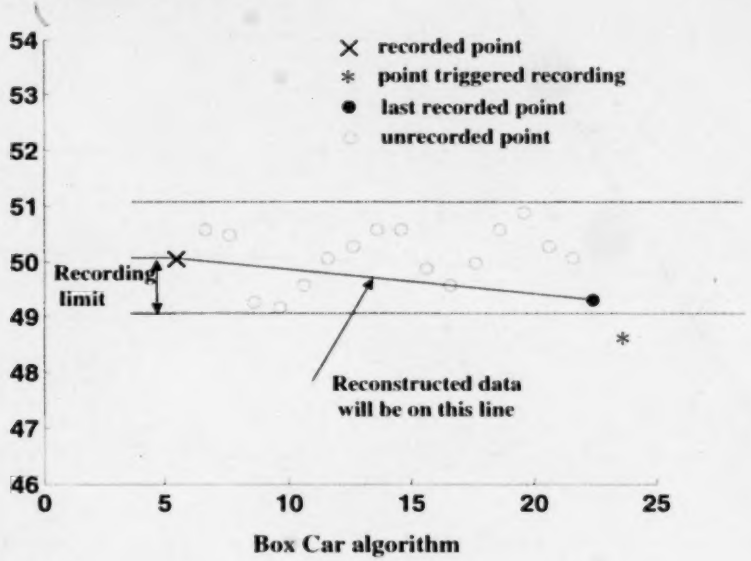


Figure 6.2: *Box car algorithm for data compression (from Hale and Sellar (1981)).*

Backward Slope (BS) Algorithm

As the name suggests, the recording limit or the compression deviation boundary lines run parallel to the projected slope calculated from the last two recorded data points. If the current data point falls outside the deviation limit boundaries, the prior point is recorded. Figure 6.3 demonstrates the method clearly. In case of a very noisy variable, the backward slope algorithm neither produces good results nor gives a higher data compression, because the slope is not a meaningful measure in such a case. Therefore, both methods were combined to give the combined Box Car and backward slope (BCBS) method.

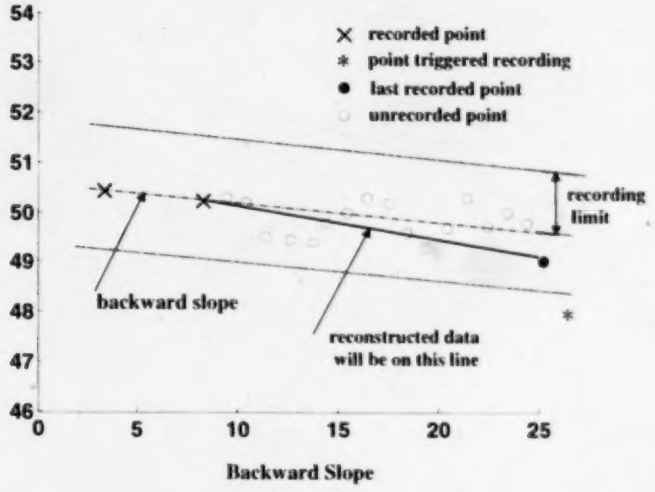


Figure 6.3: *Backward slope compression algorithm (from Hale and Sellar (1981)).*

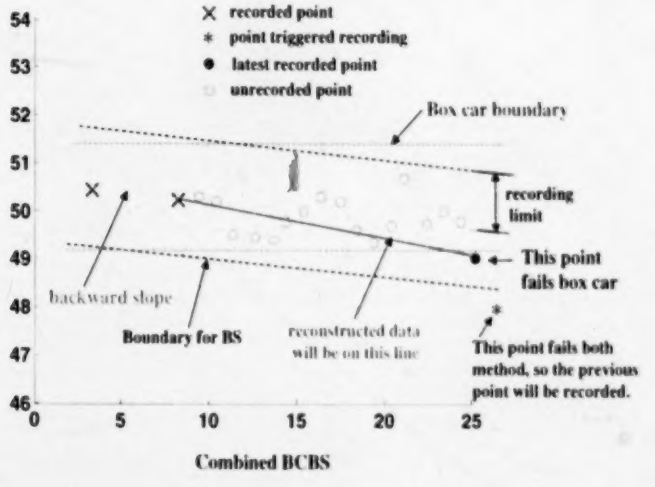


Figure 6.4: *Combined Box car and backward slope data compression algorithm (from Hale and Sellar (1981)).*

Combined Box Car and Backward Slope (BCBS) Method

In this method, the previous two methods are applied simultaneously. When the present data point fails to meet the criteria of both Box Car and Backward Slope method, the point prior to the present value is stored. See Figure 6.4 for the detailed algorithm.

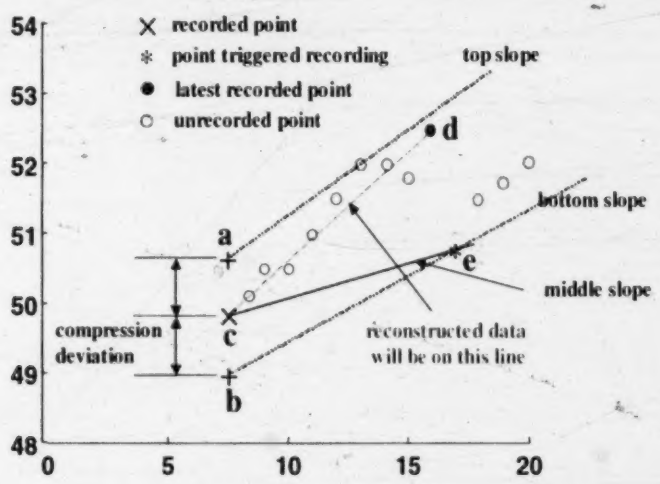


Figure 6.5: *Swinging door algorithm for data compression.*

Swinging Door Compression Algorithm

The details of this method are shown in Figure 6.5. In Figure 6.5, point 'a' is obtained from the last recorded point, 'c', plus the compression deviation, while point 'b' appears from the deduction of compression deviation from point 'c'. Points 'a' and 'b' are called pivot points. As new spot values arrive, lines are drawn from the pivot points to form a triangular envelope that includes all the spot values since 'c'. The sides of the triangle are the "doors". For instance, in Figure 6.5 all points up to point 'd' can be enveloped in a triangle. However, the next point, 'e' cannot be included in a triangle because, as shown by the dotted line, the upper and lower doors have opened wider than parallel. This signifies that a new trend started at point 'd'. Point 'd' is archived and the procedure is repeated from this point.

Issues related to direct methods: The compression factor is not specified explicitly in any of the direct methods of compression. Instead, the parameter to be set is the deviation

threshold or recording limit deviation in engineering units. Therefore, in conducting the compression tests described in this study, it was necessary to first conduct calibration trials to find the deviation thresholds corresponding to each compression factor for each data set. In the direct methods, the trends are reconstructed from the archived spot values by linear interpolation between archived points at the original sampling instants. The compression factor (CF) is defined as the ratio between the storage requirement of the original data set and that of the archived data. If the original data set had 1000 observations and 1000 time tags, a direct method with $CF = 10$ would yield 100 observations, 100 time tags and 99 linear segments.

6.3.3 Transform methods

A transform method performs an integral transformation of the original data set and the compression is performed in the transformed domain. Wavelet compression falls into this category (Donoho *et al.*, 1998). Such methods are not real-time. They require historical data since the transform is computed from an ensemble of data. Many researchers have explored wavelet compression. Nesic *et al.* (1997) demonstrated its superior performance in process data from paper making machines. Other have explored various wavelet functions selected on a case-by-case basis. Bakshi and Stephanopoulos (1996) and Misra *et al.* (2001) applied time-varying wavelet packets to achieve on-line feature extraction and noise removal from non-stationary signals. Misra *et al.* (2000) described the use of adaptive compression thresholds to control the reconstruction error. Vedam *et al.* (1998) used a multiscale representation with coarse and fine resolution linear B-splines which comprises two piece-wise linear segments. The multiscale formulation provides spline compression localization features similar to wavelet compression.

6.3.4 Selected compression method – swinging door compression

The aim of this paper is to examine the impact of data compression on activities such as minimum variance control loop benchmarking, fault detection, data reconciliation and development of inferential sensors. It concerns industrial process data and therefore focuses on piece-wise linear trending. The swinging door method was selected for detailed study as representative of industrial practice. Similar results were also observed with the BCBS and PLOT methods.

6.4 Measures of Data Quality

6.4.1 Statistical properties

Archived process data may be used for steady-state assessments such as plant production rates. Other uses include data reconciliation and mass balancing, for instance for the detection of leaks. Therefore, if compressed archived data are to be used for these purposes, the mean value of the reconstructed data should be the same as the mean of the original. The measure used was the percentage difference between the mean values (PDM) scaled by the standard deviation of the original data. The scaling allows the relative significance of any change in mean value to be assessed:

$$PDM = 100\frac{mean(y) - mean(\hat{y})}{\sigma_y} \tag{6.1}$$

where y is the uncompressed original signal, $\hat{y}$ is the reconstructed signal, and σ_y is the standard deviation of y.

Process variability has an impact on profit (Martin *et al.*, 1991; Shunta, 1995) and plant audits usually begin with a determination of the standard deviations or the variances of the time trends. Therefore, it is also necessary to determine the impact of compression on the observed variance. The measures used are the ratios between the variance of the reconstructed data ($\sigma_{\hat{y}}^2$) and the variance of the original data (σ_y^2) (*RVC*), and between σ_y^2 and the variance of the reconstruction error σ_e^2 where $e_i = y_i - \hat{y}_i$ (*RVE*). The measures are:

$$RVC = \sigma_{\hat{y}}^2 / \sigma_y^2 \tag{6.2}$$

and

$$RVE = \sigma_e^2 / \sigma_y^2 \tag{6.3}$$

If the two measures add up to 1, then the reconstruction error is the orthogonal complement of the compressed signal (i.e. the sequence $y_i - \hat{y}_i$ is uncorrelated with the sequence $\hat{y}_i$). The significance of this observation is considered in Section 6.6.2.

6.4.2 Non-linearity measure

Non-linearity assessment is starting to be used as a diagnostic tool for troubleshooting of hardware faults that may be present in the control loops (Choudhury *et al.*, 2004*d*; Choudhury *et al.*, 2002) and to make decisions about the type of model needed in inferential

sensing (Barnard *et al.*, 2001). Therefore, it is necessary to determine how the use of reconstructed data would influence non-linearity assessment.

A distinctive characteristic of a non-linear time series is the presence of phase coupling such that the phase of one frequency component is determined by the phases of others. Phase coupling leads to higher order spectral features, which can be detected in the bicoherence of a signal. The non-linearity test applied here used the bicoherence measure to assess non-linearity. The squared bicoherence is:

$$bic^2(f_1, f_2) \triangleq \frac{|B(f_1, f_2)|^2}{E[|X(f_1)X(f_2)|^2]E[|X(f_1 + f_2)|^2]} \tag{6.4}$$

where $B(f_1, f_2)$ is the bispectrum at frequencies (f_1, f_2) and is given by

$$B(f_1, f_2) \triangleq E[X(f_1)X(f_2)X^*(f_1 + f_2)], \tag{6.5}$$

$X(f_1)$ is the discrete Fourier transform of the time series $x(k)$ at the frequency f_1, $X^*(f_1)$ is the complex conjugate and E is the expectation operator. A key feature of the bispectrum is that it has a non-zero value if there is significant phase coupling in the signal x between frequency components at f_1 and f_2. The bicoherence gives the same information, but is normalized as a value between 0 and 1.

Choudhury *et al.* (2004*d*) defined two indices – the Non-Gaussianity Index (*NGI*) and the Non-Linearity Index (*NLI*) – as

$$NGI \triangleq \overline{\hat{bic}^2} - \overline{bic^2}_{crit} \tag{6.6}$$

$$NLI \triangleq |\,\hat{bic}^2{}_{max} - (\overline{\hat{bic}^2} + 2\sigma_{\hat{bic}^2})\,| \tag{6.7}$$

where $\overline{\hat{bic}^2}$ is the average squared bicoherence and $\hat{bic}^2{}_{max}$ is the maximum squared bicoherence, $\sigma_{\hat{bic}^2}$ is the standard deviation of the squared bicoherence and $\overline{bic^2}_{crit}$ is the statistical threshold/critical value obtained from the central chi-squared distribution of squared bicoherence. As outlined in (Choudhury *et al.*, 2004*d*), if both *NGI* and *NLI* are greater than zero, the signal is described as non-Gaussian and nonlinear.

6.4.3 Performance index (Harris) measures

The widely-used Harris index (Desborough and Harris, 1993) is a minimum variance benchmark of control loop performance. Significant industrial implementations of this index have been reported (Fedenczuk *et al.*, 1999; Paulonis and Cox, 2003; Desborough and Miller, 2002). It is known that the use of data compression influences the Harris

index (Thornhill *et al.*, 1999). Consequently, an issue for practitioners is to know whether compressed archived data can be used for the purposes of a minimum variance benchmark calculation.

The Harris indices for the three data sets were calculated using the method described by Desborough and Harris (1993) with an estimated time delay of 5 samples. The index is determined from the residuals between the measured controller error denoted by y and a b-step ahead prediction, $\hat{y}$.

$$r(i) = y(i) - \hat{y}(i) \tag{6.8}$$

The model for $\hat{y}$ employed 30 autoregressive terms (i.e., $m = 30$) as discussed in (Thornhill *et al.*, 1999) and in this case the prediction horizon was $b = 5$ since the time delay was estimated to be 5 sample intervals.

$$\hat{y}(i+b) = a_o + a_1 y(i) + a_2 y(i-1) + ... + a_m y(i-m+1) \tag{6.9}$$

The minimum variance benchmark is:

$$1 - \frac{\sigma_r^2}{mse(y_i^2)} \tag{6.10}$$

where σ_r^2 is the variance of the residuals r, and $mse(y_i^2)$ is the mean square value of the controller error. An index of 0 represents minimum variance control, while an index of 1 represents poor control (in which $y \approx \hat{y}$ and r is negligible). In the later case, the controller is failing to deal with predictable components such as steady offsets or a predictable oscillatory disturbance.

The reconstructed data are more predictable and thus have a worse (larger) Harris index than the original because compression removes noise and produces piecewise linear segments, that have high local predictability. Thus, there is a danger that unnecessary maintenance effort may be spent on repair of control loops wrongly identified as performing poorly.

6.5 Process Data for Compression Comparison

6.5.1 Industrial example 1

Three contrasting time series variables were chosen for the evaluation of the impact of compression, courtesy of BP. They are uncompressed liquid flow trends from continuous processes operating at steady state. Each data set comprised nearly three hours of 10s

samples representing deviations of flow in a process stream from the mean value. Figure 6.6 shows portions of the time trends while the dotted lines in Figure 6.7 show their power spectra (Welch, 1967).

Data set 1 shows a persistent oscillation characterized by an average of about 22 samples per cycle. Figure 6.7 shows that the spectrum of this signal has a broad peak at a frequency of 0.045 times the sampling frequency (i.e., 22 samples per cycle). The challenge for high fidelity compression and reconstruction is to retain the spectral peak in the frequency domain and the oscillatory features in the time domain.

Data set 2 has a tendency to stay at a constant value for a given time and then to move rapidly to a new level. These data are from a control loop which has a limit cycle caused by a sticking valve. The signal is predictable for long periods, and its spectrum shows very low frequency features because the period of oscillation is long while a series of harmonics highlights the non-sinusoidal nature of the waveform. The low frequency features and harmonics should be preserved during compression and reconstruction.

Data set 3 has little predictability, and has spectral features at all frequencies. This signal is dominated by random noise, and is from a well tuned loop operating close to minimum variance.

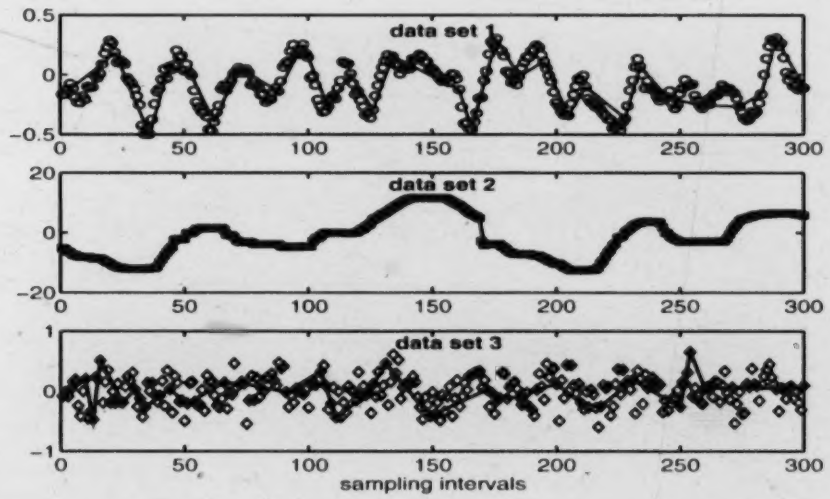


Figure 6.6: *Time trends of original data (open circles for data set 1, squares for data set 2, and diamonds for data set 3) and reconstructed data with compression factor 10 (lines).*

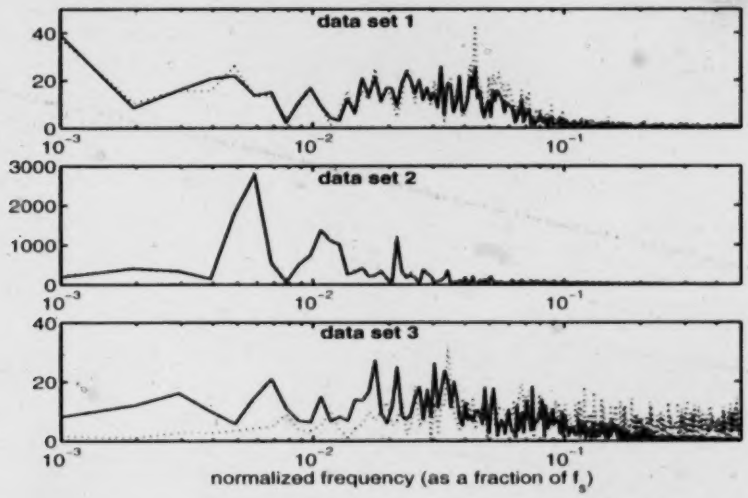


Figure 6.7: *Power spectrum of the original data and reconstructed data with compression factor 10, original spectra (dotted line) and spectra of reconstructed signals (solid line).*

6.5.2 Industrial example 2

The purpose of choosing this example is to study the effect of compression in fast and slow control loop data. Also, the results of this example will validate some of the results of Example 1. In this data set, data from three different types of control loops were chosen, courtesy of ARAMCO. They are uncompressed flow, temperature and pressure trends from continuous processes operating at steady state. Each data tag comprised nearly four hours of 15s samples representing deviations of flow in a process stream from the mean value. Figure 6.8 shows portions of the time trends, while the dotted lines in Figure 6.9 show their power spectra.

Flow loop data show a good control loop operating close to the minimum variance benchmark, and have little predictability. It has spectral features at all frequencies.

Temperature loop data have some irregular oscillations. The oscillations are not always present.

Pressure loop data have low frequency predictable features.

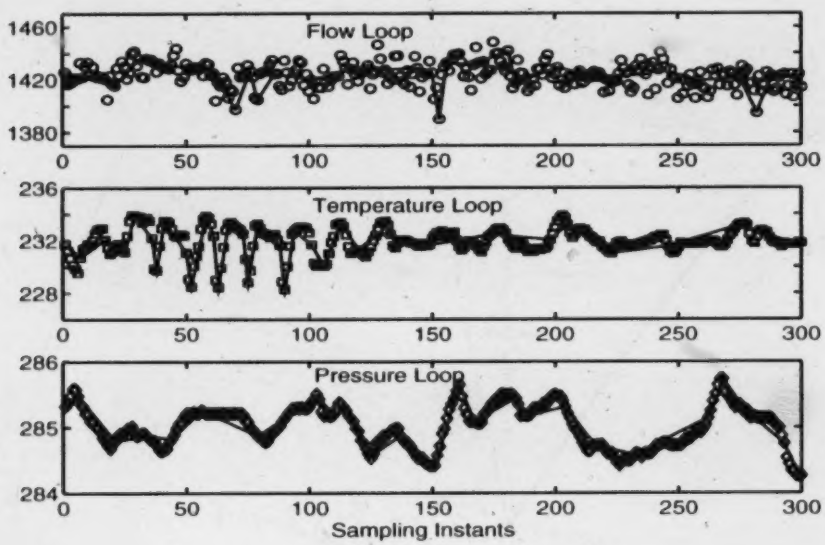


Figure 6.8: *Time trends of original data (open circles for flow loop data, squares for temperature loop data, and diamonds for pressure loop data) and reconstructed data with compression factor 10 (lines).*

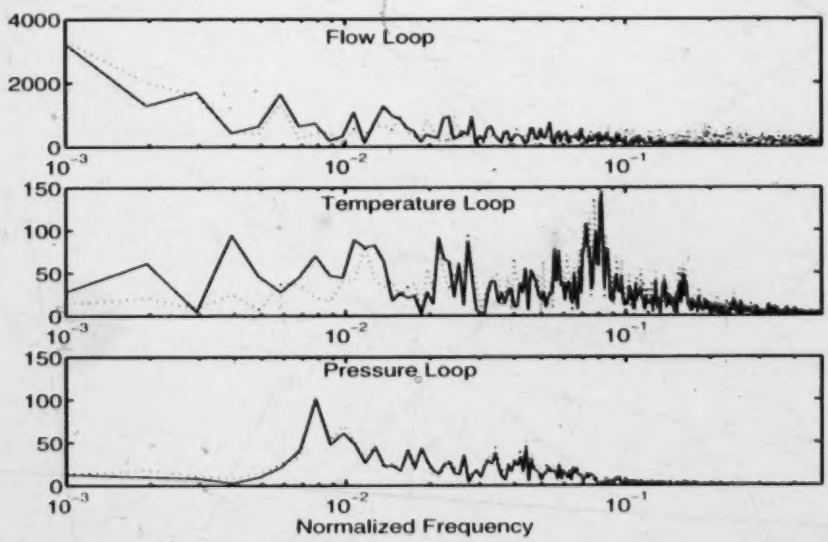


Figure 6.9: *Power spectrum of the original data and reconstructed data with compression factor 10, original spectra (dotted line) and spectra of reconstructed signals (solid line).*

6.6 Results and Discussions for Industrial Example 1

6.6.1 Visual observations

The panels in Figure 6.6 show close-up portions of the original data (as points) and reconstructions (solid line) with compression factor of 10 for data sets 1 to 3. Each complete data set had 1024 samples. Features to note include:

- Swinging door compression did not follow all the oscillations in data set 1 because with $CF = 10$, the average duration of each linear segment was longer than half of the oscillation period;
- High fidelity compression was possible with data set 2, but with data set 3 much of the randomness was lost from the reconstructed trends.

Figure 6.7 shows reconstruction in the frequency domain. The power spectra of the original signal are denoted by dotted lines and the spectra of the reconstructed signals with $CF = 10$ are shown as solid lines. When the two are not the same, a reconstruction error exists.

- The spectral feature in data set 1 at 0.045 samples per cycle was not fully captured by the reconstructed data set;
- Data sets 1 and 3 had errors at low frequency and a non-zero spectral error at $f = 0$. Therefore, the signal reconstructed after compression had a different mean value than the original;
- The low frequency harmonics of data set 2 were reproduced well, but the high frequencies of data set 3 were not captured.

The observations from the spectra reinforce and illuminate the observations from the time domain plots. The frequency domain plots also provide insight into why data set 2 is more compressible than data set 3. Data set 2 has very few spectral features, and they are at low frequency (i.e. of long duration). In contrast, data set 3 has features over the entire frequency range. Data set 2 is therefore a much simpler signal with fewer different types of behavior to capture.

6.6.2 Statistical properties

Figure 6.10 shows the behavior of the mean value and variance measures as a function of compression factor. Noteworthy observations include:

- The mean of the signal reconstructed from the archive differs from the mean of the original.
- The variances of the reconstructed data are smaller than the variance of the original signal;
- Variance measures at a given compression factor do not sum to 1.

It is concluded that data compression provides misleading information about basic statistical properties of data. Compression alters both the mean and the variance. The changes in means are only a small percentage of the standard deviation. However, the purpose of data reconciliation is often to find small shifts in the mean value that may be indicative of problems such as leaks. The shift in mean due to data compression may therefore be wrongly interpreted as evidence of a leak. Decisions of the type used in statistical process control (Wetherill and Brown, 1991) may also be erroneous if the warning and alarm limits have been based upon a statistical distribution determined from compressed archived data.

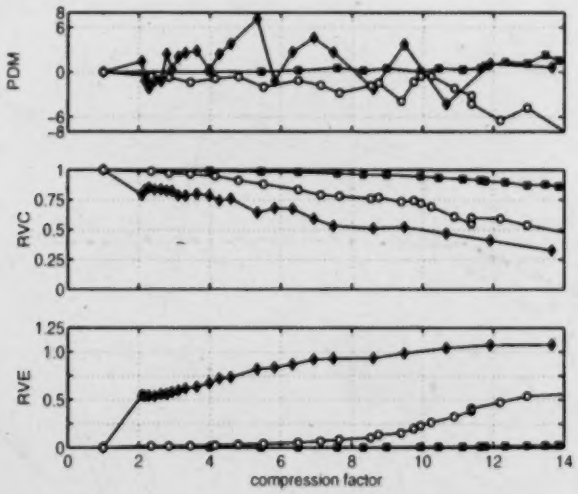


Figure 6.10: *Statistical measures as a function of CF for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).*

The sum of the measures of error variance (RVE) and compressed signal variance (RVC) was not 1. Thus, a correlation between the portion of the signal deleted during compression and the compressed signal itself exists. The implication for data-driven methods such as

inferential sensing is that some informative features have been thrown away or that some unwanted features have been retained.

6.6.3 Nonlinearity assessment

The first and second panels of Figure 6.11 show results from non-linearity assessment of the three data sets. Two of the three data sets (Data 1 and Data 3) were linear in their uncompressed state (CF=1) but became non-linear after compression and reconstruction when the compression factor exceeded 3. It demonstrates that compression induces non-linearity in the signal. Compression is a non-linear operation and the principle of superposition does not apply, i..e:

$$g(x_1(t)) + g(x_2(t)) \neq g(x_1(t) + x_2(t)) \tag{6.11}$$

and

$$g(a \times x_1(t)) \neq a \times g(x_1(t)) \tag{6.12}$$

where $x_1(t)$ and $x_2(t)$ are time domain signals, $g(x(t))$is a compressed time trend and a is a scalar factor. In the case of swinging door compression, if the signal were twice as large then the compressed signal would not merely be twice as large at the retained spot values. It would also have more piece-wise linear segments because more spot values would hit the condition for archiving.

The use of compressed archived data to assess non-linearity, for instance in an audit of control valves, may be misleading. Time may be wasted in inspection and testing of valves that are in fact operating normally.

6.6.4 Performance (Harris) index

The third row of Figure 6.11 shows the Harris index results, where 0 represents good performance close to minimum variance and 1 represents poor performance. The index increases with compression for all the data sets.

In the case of data sets 1 and 3, it is concluded that compression increases the predictability of the signal and thus affects the Harris index. Data set 2 was inherently predictable (see Section 6.5.1). The Harris index for data set 2 therefore indicated poor performance even in the uncompressed case, and did not change as much on compression as for the other two data sets.

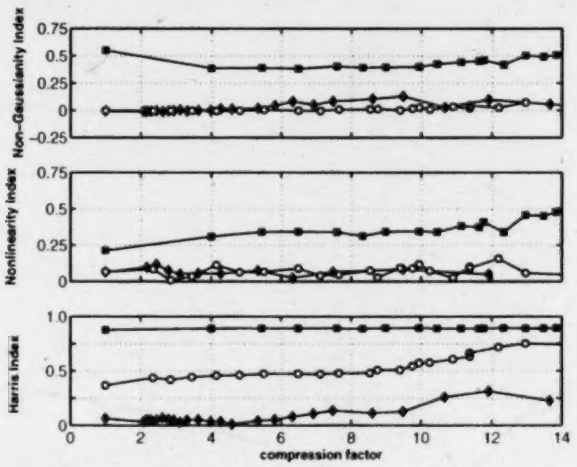


Figure 6.11: *Non-Gaussianity Index, Nonlinearity index, and Harris index as a function of compression factor for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).*

6.7 Results and Discussions for Industrial Example 2

6.7.1 Visual observations

The panels in Figure 6.8 show close-up portions of the original data (as points) and reconstructions (solid line) with compression factor of 10 for data sets 1, 2, and 3. Each data set had 1024 samples. Features to note are:

- In the reconstructed flow loop data set much of the randomness of the signal was lost.
- Oscillatory portion of the temperature trend was reconstructed well but the other portion was not.
- High fidelity compression was possible with pressure loop data because it contains mainly low frequency components.

Figure 6.9 shows reconstruction in the frequency domain. The power spectra of the original signal are shown as dotted lines, while the spectra of the reconstructed signals with $CF = 10$ are shown as solid lines. When the two are not identical, a reconstruction error exists.

- There are errors in the low and high frequency spectral features in the reconstructed flow loop data.

- Flow and temperature loop data have reconstruction errors at low frequency and a non-zero spectral error at $f = 0$. Therefore, the signal reconstructed after compression had a mean value that differed from the original (see the first panel of Figure 6.12).
- The pressure loop data was reproduced well.

6.7.2 Statistical properties

Figure 6.12 shows the behavior of the mean value and variance measures as a function of compression factor. The results observed in Example 1 also apply here. Figures 6.10 and 6.12 show similar results.

- The mean of the signal reconstructed from the archives differs from the mean of the original. However, the deviation is negligible up to a compression factor (CF) of 3.
- The variances of the reconstructed data are smaller than the variance of the original signal;
- The variance measures at a given compression factor do not sum to 1.

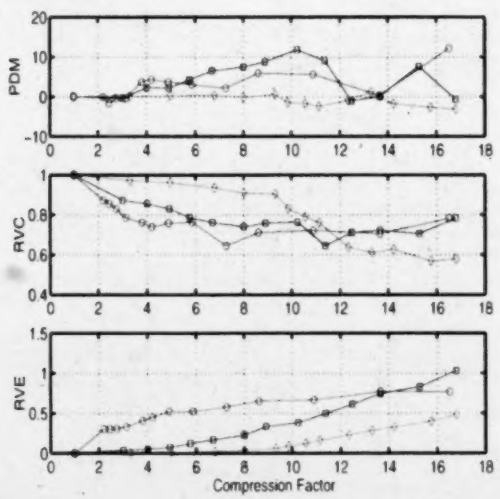


Figure 6.12: *Statistical measures as a function of CF for Flow loop (circles), temperature loop (squares) and pressure loop data (diamonds).*

6.7.3 Nonlinearity assessment

The first and second rows of Figure 6.13 depict results from non-linearity assessment of the three data sets. Compression induces non-Gaussianity and non-linearity in signals when compressed by a factor of more than 3 because uncompressed flow and pressure signals were originally Gaussian and linear. Non-Gaussianity indices increases monotonically for all three data sets, while the nonlinearity indices do not show any specific pattern – indicating the alteration of nonlinear structure of the signal with compression.

6.7.4 Performance (Harris) index

The third row of Figure 6.13 depicts the Harris index results. Again, the index increases with compression for all data sets.

In the case of flow and pressure loop data, it is concluded that compression increases the predictability of signals and thus affects the Harris index. The temperature loop shows poor performance even in the uncompressed case, and did not change as much on compression as for the other two data sets.

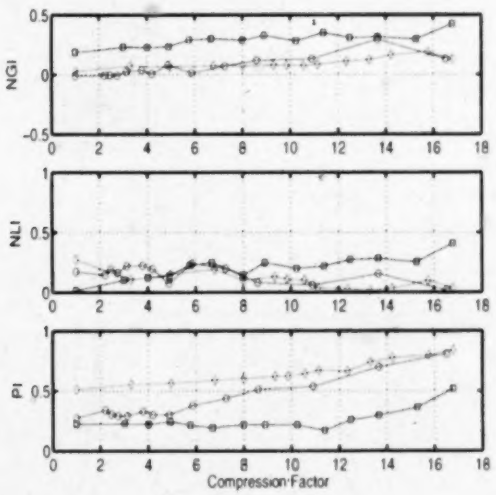


Figure 6.13: *Non-Gaussianity Index, Nonlinearity Index, and Harris index as a function of compression factor for flow loop data (circles), temperature loop data (squares) and pressure loop data (diamonds).*

6.8 Summary of Data Quality Measures

The following comments relate the results and discussion of the previous two sections to real-world industrial applications.

- Data compression changes the statistical properties of data.
- Averaging, data reconciliation and mass balancing applications should calculate the required quantities directly from the original data because data archived with swinging door compression have a different mean value after reconstruction. This could have serious implications, for example if the reconstructed data represent oil flow from an off-shore facility being monitored for taxation purposes.
- High fidelity reconstruction requires that the statistical properties of the reconstructed signal be similar to those of the original. Minimum variance and non-linearity assessment are two procedures that require high fidelity data. Swinging door compression alters these measures significantly.
- Data smoothing, feature extraction and reconstruction of events require that the events and features of interest be retained during compression. The nonorthogonality of piece-wise linear trending means that the condition is not met because the reconstruction error is correlated with the reconstructed signal. An example of a negative consequence is that the magnitude of a transient event may not be reconstructed accurately.

The performance measures for data set 3 in industrial example 1 and for flow loop in example 2 (which are random signals) were influenced by compression even at small compression factors. For instance, the RVE and RVC measures changed significantly. No random value is any more significant than any other, but the compression algorithm makes some points more significant by choosing to archive them and therefore the reconstructed signal does not have the same randomness. However, the performance measures for data sets 1 and 2 in example 1 and all loops in example 2 did not change as much for small compression factors up to about 3. For Example 1, in the third row of Figure 6.11, the results for a CF of 3 were very similar to those for the uncompressed case when CF was 1, while the second and first rows of Figure 6.11 show that non-linearity was not induced for a CF of 3 or less. For example 2, Figures 6.12 and 6.13 show that the properties did not change much until the compression factor exceeded 3. Therefore, the following heuristic rule is proposed to ensure that at least some compressed archived data can be exploited:

Data having CF ≤ 3 may be used with caution for data-driven process analyses.

It is noted, however, that certain types of process trends may allow higher compression factors because their intended use is to record constant values such as set points, targets and high and low limits.

6.9 Automated Detection of Compression

6.9.1 Motivation

The previous discussion showed that compression induces changes to many of the quantities commonly used in data-driven process analyses. However, engineers are not always in a position to examine data closely enough to detect compression because plotting and examining time trends are time-consuming. They may also not be aware of the default compression parameters set on their historians. Consequently, data-preprocessing activity usually focuses on finding and replacing bad data such as missing values and outliers. If archived data are to be used for an automated analysis, it is first necessary to test for the presence of compression.

If the number of spot values in the compressed archive and the original sampling rate are known, then the compression factor may be determined by computing the ratio between the expected number of observations and the number of archived observations. However, such information is not always available, and it may be necessary to estimate the compression factor from the reconstructed data only. An automated method for detection of piece-wise linear compression is now presented, and several guidelines are given for its application to industrial data.

6.9.2 Compression detection procedure

Because the reconstructed data set is piecewise linear, its second derivative is zero everywhere apart from at the places where the linear segments join. Therefore, the presence of the characteristic linear segments can be detected by counting zero-valued second differences $\Delta(\Delta\hat{y})$ calculated from:

$$\Delta(\Delta\hat{y})_i = \frac{(\hat{y}_{i+1} - \hat{y}_i)/h - (\hat{y}_i - \hat{y}_{i-1})/h}{h} = \frac{\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}}{h^2} \tag{6.13}$$

where $\hat{y}$ is the reconstructed signal and h is the sampling interval. The index i ranges from 2 to N-1, where N is the number of samples. Suppose the original data set had N values,

and after compression there are m archived spot values and $m-1$ linear segments. If the reconstructed data are differenced twice, there will be $n = N - m$ second differences whose values are zero. Therefore, the compression factor can be determined from:

$$CF_{est} = \frac{N}{m} = \frac{N}{N-n} \tag{6.14}$$

where $m = N - n$. For example, with 10 data points compressed to 4 archived values and 3 linear segments, there are $10 - 4 = 6$ second differences whose values are zero. Therefore, the CF is 2.5.

The method can be extended to other piecewise reconstruction methods using polynomials. For instance, if cubic spline compression were in use (Vedam *et al.*, 1998), the fourth derivatives would be zero everywhere except for at the knot points where the splines join. In that case, the compression factor would be determined from the number of fourth differences having zero values.

Figure 6.14 shows results for data sets 1, 2 and 3 in example 1 while Figure 6.15 compares the estimated compression factor with actual compression factor for example 2. Results for both examples show that the compression factor derived from counting the zero second differences was a good estimate of the true compression factor.

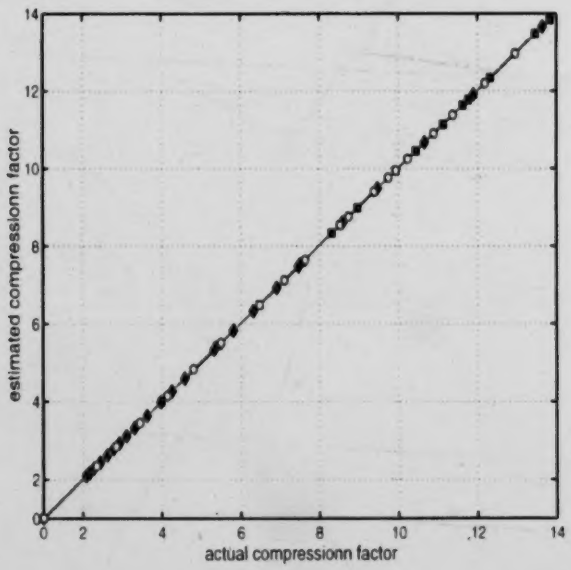


Figure 6.14: *Results from the compression estimation algorithm for data sets 1 to 3, data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).*

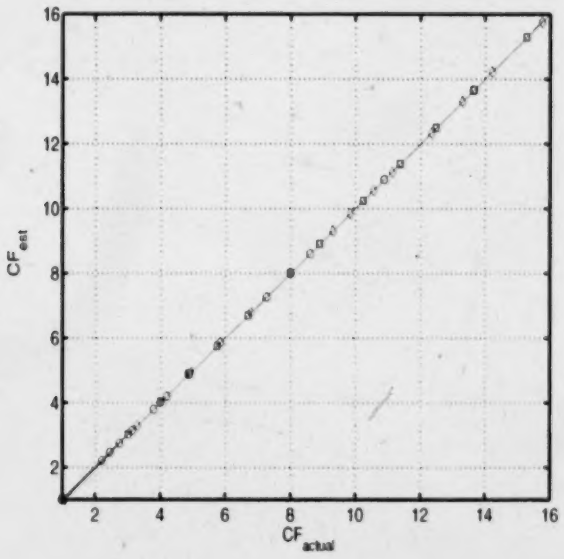


Figure 6.15: *Results from the compression estimation algorithm, flow loop (circles), temperature loop (squares) and pressure loop (diamonds).*

6.9.3 Implementation considerations

Enhancements to the basic algorithm are needed for industrial implementation for the following reasons:

- The sampling interval of the reconstructed signal may be larger than the original; e.g., the compression algorithm may have used 10s samples but the reconstruction may use 1 min samples.
- The effects of finite precision arithmetic – some computed second differences may not be exactly zero.

Suggestions for handling these cases are given here and illustrated with the industrial data of Figure 6.1.

Dealing with a larger reconstruction sampling interval

It is recommended that the compressed data be reconstructed using the same sampling interval as the original because reconstruction with a longer sampling interval leads to an underestimation of the compression factor. For example, if five 10s samples out of 120

were archived then the true compression factor is 24 ($m = 5$, $N = 120$). When the data are reconstructed using 1 minute samples, the number of piecewise linear segments does not change, but there are only 20 samples in the reconstructed data so the compression factor appears to be 4.

An effect of reconstruction with a longer sampling interval is that the true end points of the piecewise linear segments may fall between samples. Thus, x_i would be the end of one linear segment and x_{i+1} would be the start of the next, with the true end point somewhere in between. The effect on the second differences is that there are two non-zero second differences where the linear segments join, instead of one that would be expected. The presence of these pairs of non-zero second differences can be used as a warning of a sampling interval issue. If such pairs are detected, then the calculation of the compression factor has to acknowledge that each pair represents only one true archived point and the expression for the compression factor is modified to:

$$CF_{est} = \frac{N}{m/2}$$

Such pairs were detected in the industrial data of Figure 6.1. Therefore, the modified expression was used in the compression factor calculation. The estimated compression factors are shown in the right column of Figure 6.1. For instance, tag 20 has a compression factor of 41.7. It had 1428 zero second differences, 72 non-zero second derivatives in 36 pairs and 36 linear segments.

If the characteristic pairs are noticed, then a warning must be given that the compression factors have been underestimated. Figure 6.1 showed such a warning.

Finite precision arithmetic

As discussed earlier due to the effects of finite precision arithmetic, some computed second differences may not be precisely zero. This problem can be dealt with either of the following two methods.

1. The numerical values of the second differences were converted to integers. The *ceil* function in the following expressions rounds up to the next integer:

$$P = ceil\,(\log_{10}|x|)$$
$$y = x/10^P$$
$$z = y \times 10^N$$

x is the original entry in the data base having N significant figures, $P - 1$ of which are to the left of the decimal point (e.g. $P = 5$ and $N = 10$ in 1478.144165). y has the

same digits as x, but has a zero to the left of the decimal point (e.g. 0.1478144165) and z is an integer with the same digits (e.g. 1478144165). The second difference calculations were applied to the integers z.

Certain computed second differences may not be precisely zero because of arithmetic rounding errors. With the integer transformation above, the errors would expected to be ± 1. However, errors of up to ± 500 were observed. That is to say, the precision of the arithmetic used by the data historian in the reconstruction was less than 10 significant figures, although the results were reported to 10 significant figures. The following sequence illustrates the pattern of second differences observed in z for a portion of a straight line trend in tag 7 of the industrial data of Figure 6.1.

$$\left\{ \begin{array}{l} -476,\ 477,\ 0,\ -477,\ 477,\ 0,\ -1,\ -476,\ 477, \\ .\quad 0,\ 0,\ -477,\ 477,\ 1,\ 477,\ -1,\ -476,\ 477,\ 0 \end{array} \right\}$$

Any second difference in z whose absolute value was below 500 was counted for calculation of the compression factor.

If the data historian complies with a published numerical Standard (e.g., IEEE 854-1987), then the threshold for second differences may be determined from the Standard. Otherwise the threshold must be determined by observation of the arithmetic precision achieved and the number of significant figures in use, as was done in this study. There is no fundamental significance to the numerical value of ± 477 in the example presented above. The observed rounding errors appear to arise from an interplay between the original data values, the arithmetic precision and the details of the data base.

2. The second procedure will be explained with the help of an example. Consider the flow loop of industrial example 2. The second derivative of the reconstructed compressed flow trend with a compression factor of 10 resembles the plot shown in the left panel of Figure 6.16. The distribution of the second derivative (right panel of Figure 6.16) shows that all the zero second derivatives fall in the central bin, whose center is at zero. Therefore, the size of the central bin provides a good estimate of the number of zero second derivatives. The challenge in this approach is to choose the correct number of bins for the histogram. In order to solve this problem, an investigation on how the central bin size varies with the number of bins (Figure 6.17) reveals that the size of the central bin does not change much beyond a certain number of bins. This was observed for many other compressed data. Thus, the number of bins can be chosen to be greater than the number after the elbow joint in Figure

6.17. For an automatic compression detection algorithm, the number of bins for the histogram calculation can be fixed anywhere between 300 and 400 for a data set whose length is larger than 1000 samples. This was observed by the author for most cases of industrial data analyses.

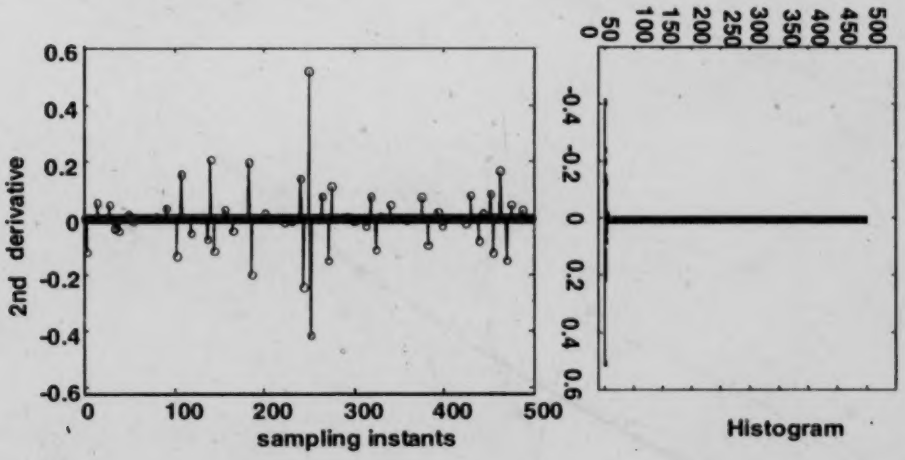


Figure 6.16: *Left panel shows the second derivative of the flow trend, Right panel shows the distribution of the second derivative.*

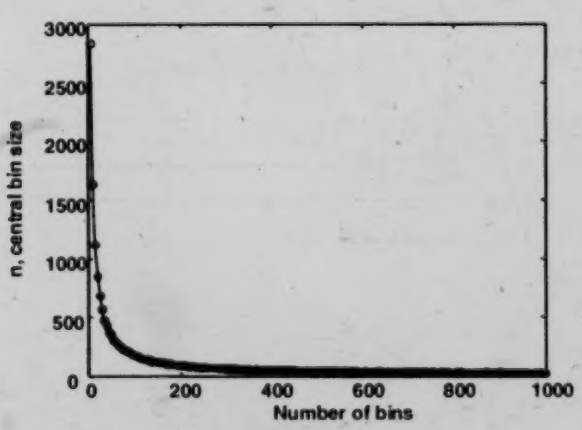


Figure 6.17: *The variation of central bin size with number of bins.*

6.10 Final Recommendations Concerning Compression

6.10.1 Discussion on the motivating example

It has been demonstrated earlier that data with $CF > 3$ are not suitable for data-driven analyses. Compression factors of up to 93.8 were present in the industrial data set of Figure 6.1 and only five tags had compression factors of three or under. It was concluded that this archived data set would not be suitable for data-driven analyses. Moreover, the algorithm issued a warning that the compression factors were underestimated. For improved estimates of compression factor, the data set should be reconstructed with the original sampling interval. The reconstruction was not attempted in this study because it was already clear that the data were much too compressed for data-driven analyses.

6.10.2 A proposal for harmless storing of data

Ideally, data should not be compressed. If compression is absolutely necessary, data should not be compressed more than a factor of 3. In order to ensure this, an algorithm such as one explained in Figure 6.18 can be implemented in data historian before storing the data.

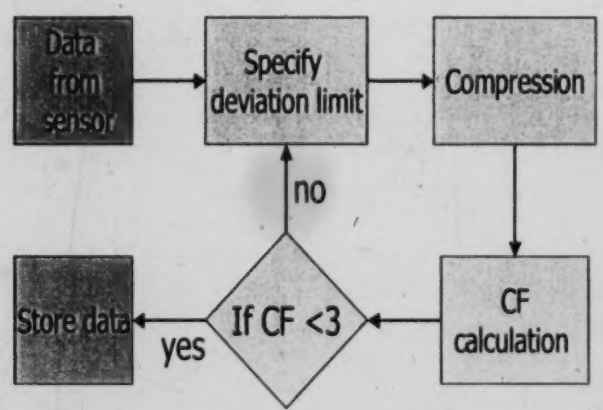


Figure 6.18: *An algorithm for storing of data with little loss of information.*

6.11 Quantization

Modern control systems work with the help of computers or digital equipments. Because computers cannot read analog signals, all analog signals must be converted to digital signals. For digitization, two steps are required: sampling and quantization. Sampling is only the first phase of acquiring data into a computer. Computational processing further requires that the samples be quantized – analog values are converted into digital form. In short, one needs to perform analog-to-digital (A/D) conversion. The number of bits of the A/D converter determines the precision of the digitized data. Old A/D converters have a lower number of bits, i.e., low resolution. Those may introduce a significant amount of quantization errors. Significant quantization errors produce oscillations (Horch, 2000) in process variables. Sometimes the quantization errors are too large to use the data for any practical analysis. A method will be described herein to quantify the amount of quantization in the process data.

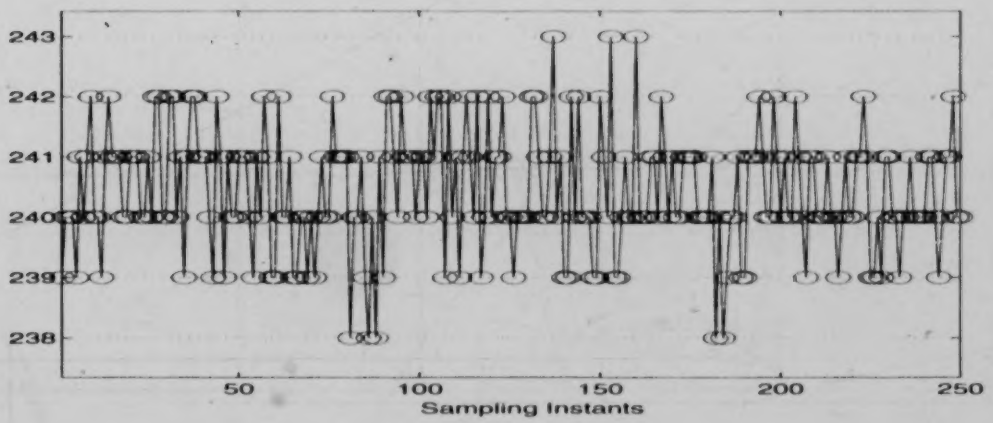


Figure 6.19: *An example of heavily quantized data representing a process variable.*

Figure 6.19 shows an example of a heavily quantized process data. The quantized data usually resembles staircases. They are easy to be detected through visualization of the data. However, to check quantization in hundreds of variables in the data preprocessing stage, an automatic method of detecting quantization is required. A quantization factor has been defined as

$$QF = \frac{Q_{level}}{\sigma} \tag{6.15}$$

where Q_{level} is the quantization level present in the data and can be obtained from the

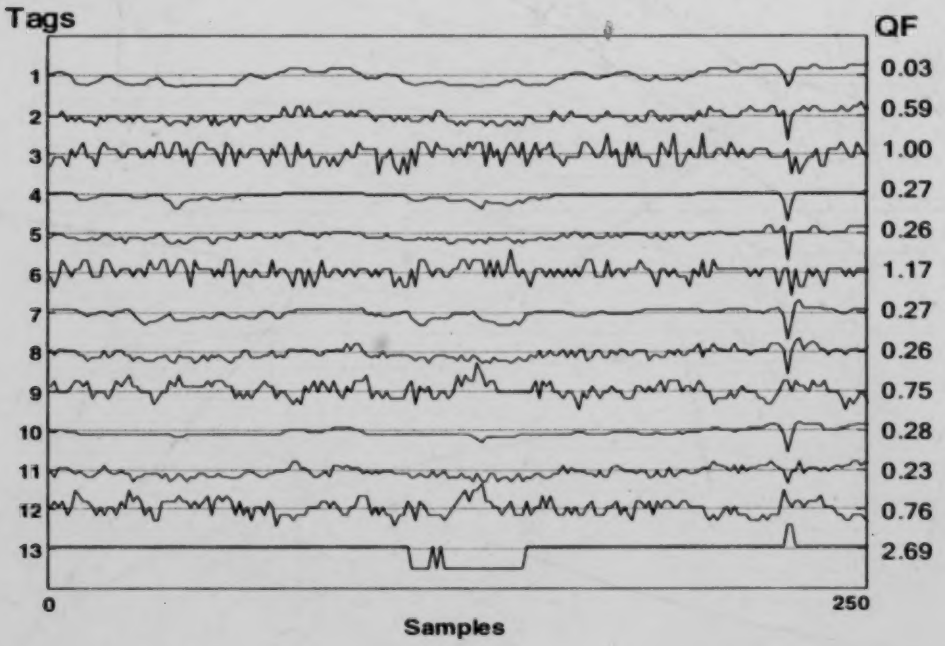


Figure 6.20: *An industrial data set where some of the variables were heavily quantized.*

minimum value of the non-zero first difference of the data, σ is the standard deviation of the data, and QF is the quantization factor. For data that have no quantization problems, the magnitude of QF should be close to zero. In contrast, for a heavily quantized data set the magnitude of QF will be closer to 1. Figure 6.20 depicts an industrial example of heavily quantized data. This data set represents vibration monitoring data of a large pump. The right side of the figure shows the quantization factor estimated for each data tag. Clearly, tags 3, 6, 9, 12, and 13 are heavily quantized. The example shown in Figure 6.19 represents tag 3 of this industrial data set for which the estimated quantization factor was 1.0, indicating heavy quantization of the data.

For most plants that implement modern control systems, quantization is not a problem because they use high resolution A/D converters. However, many plants still use old control systems and sensors, where quantization remains a problem. This automatic quantization detection algorithm can be used in the data preprocessing stage to check for quantization, and may serve as a useful diagnostic tool to decide which variables are not worth further analysis.

6.12 Conclusions

Time and frequency domain plots were presented for data from continuous processes to show how well the trends were reconstructed after compression. Piecewise linear compression using the swinging door algorithm altered key statistical features of the data set such as mean and standard deviation. Other data-driven analyses were also influenced.

A procedure for detection of compression during the pre-processing stage of a data-driven analysis, plus additional features required for its application to industrial data were presented. An expression based upon estimation of the number of zero-value second derivatives provided a lower bound for the compression factor. To reconstruct the data at the original sampling interval is important for an accurate assessment of the compression factor. If the reconstruction interval is longer than the original, then characteristic pairs of non-zero second derivatives become evident. In such an event, a warning must be issued to reconstruct the signal at the correct sampling rate.

On the basis of this study, it is strongly recommended that caution be exercised in the use of compression in process data archives. Because of pressure from customers and the cheaper costs of storage, newer data historians (e.g. AspenWatch) are beginning to use uncompressed data. It is hoped that this study will provide end-users, who wish to eliminate the use of data compression, with solid, quantitative reasons for doing so.

Finally, the quantization problem in process data analysis has been discussed. Quantization may cause limit cycles in control loops. An automatic method to detect and quantify quantization has been presented.

7

Conclusions

This chapter lists the contributions of this thesis and the directions identified for future research.

7.1 Contributions of This Thesis

The main contributions of this thesis can be summarized as following:

- A tutorial introduction of a complex subject matter, higher order statistics, has been presented in Chapter 2. Unlike most discussions in the literature, this chapter is readily understood and self-explanatory due to the illustrative examples.
- The use of higher order moments is suggested for detection of of nonlinearities in a signal. Two new indices – the Non-Gaussianity Index (NGI) and the Nonlinearity Index (NLI) – have been developed for detecting nonlinearities in time series.

- An algorithm for diagnosing the causes of poor performance of control loops has been developed and applied successfully to industrial and simulation data sets.
- Much confusion surrounding the term 'stiction' exists in the literature. Through a detail model of stiction and a thorough discussions of all terms related to stiction, stiction has been clearly explained and defined in this thesis. A new generalized formal definition of 'stiction' has been proposed.
- A two-parameter data driven model has been developed for stiction. The model is easy to understand, relatively simple to use, and yet powerful enough to simulate stiction phenomena in control valves.
- An automatic algorithm for detection and quantification of stiction in control valves has been developed and successfully evaluated through several industrial case studies.
- The impact of compression on data quality has been systematically studied and presented. It is recommended that data should not be compressed unless storage space is extremely limited.
- An automatic method for detection and quantification of compression in process data has been developed. A new method for storing data with little loss of information has been recommended.
- An automatic method for detecting quantization in process data has been developed.

7.2 Recommendations for Further Work

Diagnosis of the causes of poor control loop performance remains a challenging area for further research. In particular, an automatic procedure with minimum human intervention must be developed. The procedure must be non-invasive, plant-friendly, easy to understand, and simple to implement. Further work in this and related areas should consider all these issues, and can be categorized as follows:

- Most disturbances entering a control loop are assumed to be linear in the diagnosis of poor loop performance. Further investigations should be carried out to examine the effect of nonlinear disturbances on the loop diagnosis results.

- The quantified stiction in this thesis is termed as 'apparent stiction' because of the effect of loop dynamics, especially the controller action that compensates for some stiction. The effect of controller and loop dynamics on the quantified stiction should be investigated further.

- Because of the absence of a by-pass line, a sticky valve cannot be taken out of service for maintenance immediately. A method that can facilitate continued operation of the plant until the next planned shutdown – minimizing the impact of sticky valves – would be invaluable for the process industry.

- The method and algorithm for the diagnosis of poor loop performance has been developed based on the assumption of a single loop. It does not take into account the multivariate nature of the process. Further research should be done to investigate the plant-wide propagative nature of disturbances or faults.

- An automatic procedure for the root-cause diagnosis of plant-wide oscillation(s) would be useful to process industries. A detailed study aimed at understanding the mechanism(s) of oscillation propagation, and amplification and attenuation of oscillations as they travel through various process units, should be performed.

- Process model(s) identified from compressed data is inadequate. A systematic study should be performed to quantify the effect of data compression on model identification.

- Alternative methods of data compression and reconstruction that do not cause significant degradation of data quality need to be investigated.

Bibliography

Abramowitz, M. and I. A. Stegun (1972). *Handbook of Mathematical Functions*. Dover Publications. New York.

Agrawal, P. and S. Lakshminarayanan (2003). Tuning pid controllers using achievable performance indices. *Ind. Eng. Chem. Res.* **42**, 5576–5582.

Amin, J., B. Friedland and A. Harnoy (1997). Implementation of a friction estimation and compensation technique. *IEEE control systems* p. 7176.

Armstrong-Hélouvry, B., P. Dupont and C. C. De Wit (1994). A survey of models, analysis tools and compensation methods for the control of machines with friction. *Automatica* **30**(7), 1083–1138.

Aspentech (2001). Analysis of data storage technologies for the management of real-time process manufacturing data. Accessed April 2, 2003, from http:// www.advanced-energy.com/Upload/ symphony_wp_infoplus.pdf.

Aubrun, C., M. Robert and T. Cecchin (1995). Fault detection in control loops. *Control Engineering Practice* **3**, 1441–1446.

Bakshi, B. and G. Stephanopoulos (1996). Compression of chemical process data through functional approximation and feature extraction. *AIChE Journal* **42**, 477–492.

Barnard, J.P., C. Aldrich and M. Gerber (2001). Identification of dynamic process systems with surrogate data methods. *AIChE Journal* **47**, 2064–2075.

Baumann, H. D. (1994). *Control Valve Primer: A user's guide*. 2 ed.. Instrument Society of America. NC, USA.

Bergstrom, J. and G. A. Dumont (2003). Adaptive control of valves with friction. *Pulp and Paper Canada* **104**(6), 23–27.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum. New York.

Bialkowski, W. L. (1992). Dreams vs. reality: A view from both sides of the gap. In: *Control Systems '92*. Whistler, BC, Canada. pp. 283–294.

Brillinger, D. R. and M. Rosenblatt (1967*a*). Asymptotic theory of estimates of k^{th} order spectra. *Spectral Analysis of Time Series* pp. 153–188. B. Harris, ed.

Brillinger, D. R. and M. Rosenblatt (1967*b*). Computation and interpretation of k^{th} order spectra. *in Spectral Analysis of Time Series* pp. 189–232. B. Harris, ed.

Bristol, E.H. (1990). Swinging door trending: adaptive trend recording. In: *Proceedings of the ISA National Conference*. pp. 749–753.

Chandaran, V. and S. L. Elgar (1991). Mean and variance of estimates of the bispectrum of a harmonic random process – an analysis including leakage effects. *IEEE transactions on Signal Processing* **39**, 2640–2651.

Choudhury, M. A. A. S. and S. L. Shah (2001). Detection and diagnosis of system nonlinearities using higher order statistics. Internal Report IRC021030. Department of Chemical and Materials Engineering, University of Alberta, Canada.

Choudhury, M. A. A. S., N. F. Thornhill and S. L. Shah (2004*a*). A data-driven model for valve stiction. In: *The proceedings of ADCHEM 2003, Jan 11-14, 2004*. Hong Kong.

Choudhury, M. A. A. S., N. F. Thornhill and S. L. Shah (2004*b*). Modelling valve stiction. *Control Engineering Practice, In Press*.

Choudhury, M. A. A. S., S. L. Shah and N. F. Thornhill (2002), Detection and diagnosis of system nonlinearities using higher order statistics. In: *The proceedings of the 15th IFAC World Congress, July 2002*. Barcelona, Spain.

Choudhury, M. A. A. S., S. L. Shah and N. F. Thornhill (2004*c*). Detection and quantification of control valve stiction. In: *The proceedings of DYCOPS 2004, July 5-7, 2004*. Cambridge, USA.

Choudhury, M. A. A. S., S. L. Shah and N. F. Thornhill (2004*d*). Diagnosis of poor control loop performance using higher order statistics. *Automatica* **40**(10), 1719–1728.

Collis, W.B., P.R. White and J.K. Hammond (1998). Higher-order spectra: The bispectrum and trispectrum. *Mechanical Systems and Signal Processing* **12**, 375–394.

Cook, P. A. (1986). *Nonlinear Dynamical Systems*. Prentice-Hall International. London.

Crowe, J.A., N.M. Gibson, M.S. Woolfson and M.G. Somekh (1992). Wavelet transform as a potential tool for ecg analysis and compression. *Journal Biomedical Engineering* **14**, 268–272.

Desborough, L. and R. Miller (2002). Increasing customer value of industrial control performance monitoring - Honeywell's experience. In: *AIChE Symposium Series 2001*. number 326. pp. 172–192.

Desborough, L. and T. Harris (1993). Performance assessment measures for univariate feedforward / feedback control. *The Canadian Journal of Chemical Engineering* **71**, 605–616.

Desborough, L., R. Miller and P. Nordh (2000). Regulatory control survey. Honeywell, Unpublished manuscript.

Donoho, D.L., M. Vetterli, R.A. DeVore and I. Daubechies (1998). Data compression and harmonic analysis. *IEEE Transactions on Information Theory* **44**, 2435–2476.

Dulyakarn, P. and Y. Rangsanseri (2001). Fuzzy c-means clustering using spatial information with application to remote sensing. Presented at the 22nd Asian Conference on Remote Sensing, November 5-9, 2001, Singapore.

Elgar, S. and R. T. Guza (1988). Statistics of bicoherence. *IEEE Transactions on Acoustics, Speech and Signal Processing* **36**, 1667–1668.

Ender, D. (1993). Process control performance: Not as good as you think. *Control Engineering* **40**, 180–190.

EnTech (1998). *EnTech Control Valve Dynamic Specification (version 3.0)*.

Fackrell, J. W. A. (1996). Bispectral Analysis of Speech Signals. PhD thesis. Department of Electrical Engineering, The University of Edinburgh. Edinburgh, UK.

Fedenczuk, P., P. Fountain and R. Miller (1999). Loop scout, rpid and profit controller team up to produce significant benefits for bp. Accessed Apr 2nd 2003, from http://loopscout.com/loopscout/info/bpamoco.pdf.

Fisher-Rosemount (1999). *Control Valve Handbook*. Fisher Controls International, Inc. Marshalltown, Iowa, USA.

Fitzgerald, B. (1995). *Control Valve for the Chemical Process Industries*. McGraw-Hill, Inc. New York.

Gerry, J. and M. Ruel (2001). How to measure and combat valve stiction online. Instrumetnation, Systems and Automated Society. Houston, Texas, USA. http://www.expertune.com/articles/isa2001/StictionMR.htm.

Hagglund, T. (1995). A control loop performance monitor. *Control Engineering Practice* **3**(11), 1543–1551.

Hagglund, T. (2002). A friction compensator for pneumatic control valves. *Journal of Process Control* **12**(8), 897–904.

Hale, J. C. and H. L. Sellars (1981). Historical data recording for process computers. *Chemical Engineering Progress* **77**, 38–43.

Harris, T. J. (1989). Assessment of control loop performance. *The Canadian Journal of Chemical Engineering* **67**, 856–861.

Hatipoglu, C. and U. Ozguner (1998). Robust control of systems involving non-smooth nonlinearities using modifed sliding manifolds. In: *American control conference*. Philadelphia, PA. pp. 2133–2137.

Hinich, M. J. (1982). Testing for gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis* **3**, 169–176.

Horch, A. (1999). A simple method for detection of stiction in control valves. *Control Engineering Practice* **7**, 1221–1231.

Horch, A. (2000). Condition Monitoring of Control Loops. PhD thesis. Department of Signals, Sensors and Systems, Royal Institute of Technology. Stockholm, Sweden.

Horch, A., A. J. Isaksson and K. Forsman (2000). Diagnosis and characterization of oscillations in process control loops. In: *Proceedings of the Control Systems 2000*. Victoria, Canada. pp. 161–165.

Horch, A. and A. J. Isaksson (1998). A method for detection of stiction in control valves. In: *Proceedings of the IFAC workshop on On line Fault Detection and Supervision in the Chemical Process Industry*. Session 4B. Lyon, France.

Ifeachor, E. C. and B. W. Jervis (1993). *Digital Signal Processing: A Practical Approach*. Addison-Wesley.

ISA Committee SP51 (1996). Method of evaluating the performance of positioners with analog input signals and pneumatic output. Technical Report ANSI/ISA-75.13-1996. Instrument Society of America.

ISA Committee SP51 (2001). Control valve terminology. Technical Report ANSI/ISA-75.05.01-2000. Instrument Society of America.

ISA Subcommittee SP75.05 (1979). Process instrumentation terminology. Technical Report ANSI/ISA-S51.1-1979. Instrument Society of America.

Johnson, R. A. and D. W. Wichern (1998). *Applied Multivariate Statistical Analysis*. Prentice-Hall. New Jersey.

Kano, M., H. Maruta, H. Kugemoto and K. Shimizu (2004). Practical model and detection algorithm for valve stiction. In: *The Proceedings of the 7th DYCOPS, July 5-7, 2004*. Boston, USA.

Karczewicz, M. and M. Gabbouj (1997). Ecg data compression by spline approximation. *Signal Processing* **59**, 43–59.

Karnopp, D. (1985). Computer simulation of stic-slip friction in mechanical dynamical systems. *Journal of Dynamic Systems, Measurement, and Control* **107**, 420–436.

Kayihan, A. and F. J. Doyle III (2000). Friction compensation for a process control valve. *Control Engineering Practice* **8**, 799–812.

Kennedy, J.P. (1993). Data treatment and applications - future of the desktop. In: *Proceedings of FOCAPO*. CACHE.

Kim, Y. C. and E. J. Powers (1979). Digital bispectral analysis and its applications to nonlinear wave interactions. *IEEE Transactions on Plasma Science* **PS-7**, 120–131.

Mah, R.S.H., A. C. Tamhane, S. H. Tung and A.N. Patel (1995). Process trending with piecewise linear smoothing. *Comput. Chem. Eng* **19**, 129–137.

Martin, G. D., L.E. Turpin and R. P. Cline (1991). Estimating control function benefits. *Hydrocarbon Processing* pp. 68–73.

McMillan, G. K. (1995). Improve control valve response. *Chemical Engineering Progress* **91**, 77–84.

Mendel, J. M. (1991). Tutorial on higher order statistics (spectra) in signal processing and systems theory: Theoretical results and some applications. *Proceedings of the IEEE* **79**, 278–305.

Miao, T. and D. E. Seborg (1999). Automatic detection of excessively oscillatory control loops. In: *Proceedings of the 1999 IEEE International Conference on Control Applications*. Kohala Coast-Island of Hawai'i, Hawai'i, USA.

Misra, M., S. Kumar, S. J. Qin and D. Seeman (2000). On-line data compression and error analysis using wavelet technology. *AIChE Journal* **46**, 119–132.

Misra, M., S. Kumar, S. J. Qin and D. Seeman (2001). Error based criterion for on-line wavelet data compression. *Journal of Process Control* **11**, 717–731.

Nesic, Z., M. Davies and G. Dumont (1997). Paper machine data analysis and compression using wavelets. *TAPPI Journal* **80**, 191–203.

Nikias, C. L. and A. P. Petropulu (1993). *Higher-Order Spectra: A nonlinear signal processing frame work*. Prentice-Hall. NJ, USA.

Nikias, C. L. and J. M. Mendel (1993). Signal processing with higher order spectra. *IEEE Signal Processing Magazine* pp. 10–37.

Nikias, C. L. and M. R. Raghuveer (1987). Bispectral estimation: A digital signal processing framework. *Proceedings of the IEEE* **75**, 869–891.

Olsson, H. (1996). Control Systems with Friction. PhD thesis. Department of automatic Control, Lund Institute of Technology. Sweden.

OSI Software Inc. (2002). Pi data storage component overview. Accessed April 2, 2003, from http:// www.osisoft.com/ 270.htm.

Paulonis, M. A. and J. W. Cox (2003). A practical approach for large-scale controller performance assessment, diagnosis, and improvement. *Journal of Process Control* **13**(2), 155–168.

Piipponen, J. (1996). Controlling processes with nonideal valves: Tuning of loops and selection of valves. In: *Control Systems, '96*. Nova Scotia, Canada. pp. 179–186.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling (1986). *Numerical Recipes*. Cambridge University Press. Cambridge.

Qin, S. J. (1998). Control performance monitoring – a review and assessment. *Comput. Chem. Eng.* **23**, 173–186.

Rao, T. S. and M. M. Gabr (1980). A test for linearity and stationarity of time series. *Journal of Time Series Analysis* **1**(1), 145–158.

Rao, T. S. and M. M. Gabr (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Vol. 24. Springer-Verlag. New York.

Rengaswamy, R., T. Hagglund and V. Venkatasubramanian (2001). A qualitative shape analysis formalism for monitoring control loop performance. *Engng. Appl. Artificial Intell.* **14**, 23–33.

Riggs, J. B. (1999). *Chemical Process Control*. Ferret Publishing. Texas, USA.

Rosenblatt, M. and J. W. Van Ness (1965). Estimation of the bispectrum. *Ann. Math. Stat.* pp. 420–436.

Ruel, M. (2000). Stiction: The hidden menace. *Control Magazine*. http://www.expertune.com/articles/RuelNov2000/stiction.html.

Sharif, M. A. and R. I. Grosvenor (1998). Process plant condition monitoring and fault diagnosis. *Proc. Instn. Mec. Engrs.* **212**(Part E), 13–30.

Shinskey, F. G. (1990). How good are our controllers in absolute performance and robustness. *Measurement and Control* **23**, 114–120.

Shunta, J. P. (1995). *Achieving world class manufacturing through process control*. Prentice-Hall. NJ, USA.

Smith, S. W. (1998). *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing. CA, USA.

Stanfelj, N., T. E. Marlin and J. F. MacGregor (1993). Monitoring and diagnosing process control performance: The single-loop case. *Ind. Eng. Chem. Res.* **32**, 301–314.

Stenman, A., F. Gustafsson and K. Forsman (2003). A segmentation-based method for detection of stiction in control valves. *International Journal of Adaptive Control and Signal Processing* **17**, 625–634.

Stuart, A. and J. K. Ord (1987). *Kendall's Advanced Theory of Statistics*. Vol. 1. 5^th ed.. Charles Griffin and Company Limited. London.

Swami, A., J. M. Mendel and C. L. Nikias (1993). *Higher-Order Spectral Analysis Toolbox Users Guide*. United Signals & Systems Inc. CA, USA.

Taha, O., G. A. Dumont and M. S. Davies (1996). Detection and diagnosis of oscillations in control loops. In: *Proceedings of the* 35^{th} *conference on Decision and Control*. Kobe, Japan. pp. 2432–2437.

Tao, G., S. Chen and S. m. Joshi (2002). An adaptive control scheme for systems with unknown actuator failures. *Automatica* **38**, 1027–1034.

Terdik, Gy. and J. Máth (1998). A new test of linearity of time series based on the bispectrum. *Journal of Time Series Analysis* **19**(6), 737–753.

Thornhill, N. F., B. Huang and H. Zhang (2003*a*). Detection of multiple oscillations in control loops. *Journal of Process Control* **13**, 91–100.

Thornhill, N. F., F. Cox and M. Paulonis (2003*b*). Diagnosis of plant-wide oscillation through data-driven analysis and process understanding. *Control Engineering Practice* **11**(12), 1481–1490.

Thornhill, N. F., M. A. A. S. Choudhury and S. L. Shah (2004). The impact of compression on data-driven process analyses. *Journal of Process Control* **14**, 389–398.

Thornhill, N. F., M. Oettinger and P. Fedenczuk (1999). Refinery-wide control loop performance assessment. *Journal of Process Control* **9**, 109–124.

Vedam, H., V. Venkatasubramanian and B. Bhalodia (1998). A b-spline based method for data compression, process monitoring and diagnosis. *Comput. Chem. Eng.* **22**, S827–S830.

Wallén, A. (1997). Valve diagnostics and automatic tuning. In: *Proceedings of the American Control Conference*. Albuquerque, New Mexico. pp. 2930–2934.

Watson, A.B.(ed) (1993). *Digital images and human vision*. MIT Press. Cambridge, USA.

Watson, M.J., A. Liakopoulos, D. Brzakovic and C. Georgakis (1998). A practical assessment of process data compression techniques. *Ind. Eng. Chem. Res* **37**, 267–274.

Welch, P.D. (1967). The use of fast fourier transforms for the estimation of power spectra. *IEEE Trans Audio & Electroacoustics* **AU-15**, 70–73.

Wetherill, G.B. and D.W. Brown (1991). *Statistical Process Control*. Chapman and Hall. London, UK.

Whalen, B. R. (1983). *Basic Instrumentation*. 3rd ed.. Petroleum Extension Service (PETEX). Austin, TX.

Yuan, J. (1999). Testing linearity for stationary time series using the sample interquartile range. *Journal of Time Series Analysis* **21**, 713–722.